2014 update

BITS training: How to get the job done!

Hands-On Series

Analysis of public microarray datasets

FOR

VIB'IES

Learn to:

- *Find relevant GEO datasets*
- *Analyze data with GEO2R*
- *Try GDS clustering tools*
- *Use other public tools*
- *Find links to Functions, Diseases & Pathways*

powered by:

NCBI

©VIB-BITS, 2013-2015

BITS VIB
BIOINFORMATICS TRAINING AND SERVICE FACILITY

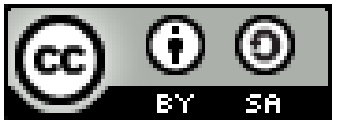
VIB

This work is owned by BITS, the bioinformatics facility of the VIB (http://www.vib.be)

This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License

You are free to:
Share — to copy, distribute and transmit the work
Remix — to adapt the work

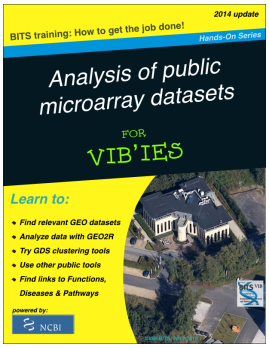
Under the following conditions:
Attribution — You must attribute the work referring to the original author and licensor (VIB)
(but not in any way that suggests that they endorse you or your use of the work)
Share Alike — If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.



Hands-on Analysis of public microarray datasets

From BioWareWIKI

"Date: October, 17 2014, from 9h30 to 17h00"



[Main_Page]

Contents

- 1 Introduction
  - 1.1 Summary
  - 1.2 Required skills
  - 1.3 Morning Session
  - 1.4 Afternoon Session
  - 1.5 More Info
- 2 Exercises
- 3 Additional resources
  - 3.1 Additional tutorials
  - 3.2 Web-services and resources
  - 3.3 Meta Analysis Resources
  - 3.4 Commercial resources licensed by VIB
  - 3.5 Do you still need MORE?

Introduction

Summary

This basic training will give you an overview of the what **GEO**[1] has to offer. Several experiments will be analyzed using simple tools to obtain differential gene lists. An introduction to downstream tools dedicated to functional enrichment will close the session.

Required skills

This training is meant for biologists with little or no data of their own that need to identify genes of interest associated to a given biological problem. The participants do not need any prior knowledge of programing.

Morning Session

- Find relevant data on GEO
- Analyze using the NCBI GEO2R utility
- Continue the GEO2R analysis in RStudio (intro)
- Find cluster using the NCBI GEO DataSets browser
- Analyze the same data using RobiNA
- Perform functional enrichment on the DE gene-lists using public tools

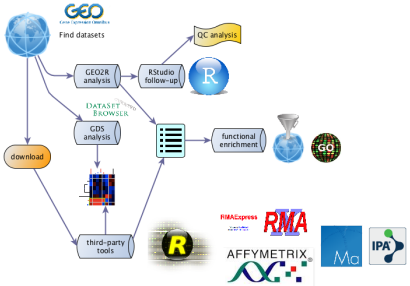
Afternoon Session

- Users search GEO datasets and analyze them with tools discovered during the morning session
- Users with access to CLC Main can follow the CLC tutorial **(VIB-only)**
- Users with access to IPA can follow the IPA tutorial **(VIB-only)**

More Info

- More on the VIB website [2]
- Related VIB training sessions [3]
- Related BITS Website pages [4]

Exercises



You will find in this section exercises performed during the hand-on session.

- **PubMA_Exercise.1** Search **GEO** to find public datasets related to one's project
- **PubMA_Exercise.2** Compute differential analysis using **GEO2R** within the NCBI web-portal (and follow-up in **RStudio**)
  - **PubMA_Exercise.2b Optional** follow-up analysis demo in **RStudio**
- **PubMA_Exercise.3** Clustering using the **GEO Dataset browser** (only for data with attached GDS ID)
- **PubMA_Exercise.4** Full **RobiNA** analysis as a standalone desktop alternative to GEO2R and Bioconductor
- **PubMA_Exercise.5 Web-tools** for functional enrichment of the obtained lists to identify key biological functions

---

- **Analyze_GEO_data_with_the_Affymetrix_software Optional** exercise using the Affymetrix **Transcription Analysis Console** (Windows only - free program)
- **Normalize CEL files with RMAExpress** (Windows and MacOSX - free program)
- **PubMA_Exercise.6 Optional** exercise using the fully integrated **CLC Main workbench** (**for VIB users and CLC license owners**)
- **PubMA_Exercise.7 Optional** IPA analysis of the GEO2R DE table (**for VIB users and IPA license owners**)

Additional resources

Additional tutorials

- **Find Transcriptome Signatures with TranscriptomeBrowser** *Alt* option to get enrichment from the GEO data (and much more)
- **Analyze your own microarray data in R/Bioconductor** See how to analyze your own microarray data in R/Bioconductor

Web-services and resources

Only few of the following resources will be used during this training.

- **GEO2R**: www.ncbi.nlm.nih.gov/geo/geo2r/
- **The Broad Institute (http://www.broadinstitute.org/)** has developped a number of tools including **GenePattern**[5] **Gene-E** [6] without forgetting the famous **GSEA** platform [7].
- **DAVID**: http://david.abcc.ncifcrf.gov (BITS WIKI: http://wiki.bits.vib.be/index.php/Exercises_on_Gene_Ontology)
- **Enrichr**: http://amp.pharm.mssm.edu/Enrichr/
- **webgestalt**: http://bioinfo.vanderbilt.edu/webgestalt/
- **TranscriptomeBrowser** [8] works as standalone on your computer and look very impressive when in good hands. Please follow the **webcast (http://tagc.univ-mrs.fr/tbrowser/index.php?option=com_content&task=view&id=35&Itemid=28)**.

Please feel free to discover these other ones with more Plant-dedicated resoures than above

- **PlantGSEA** http://structuralbiology.cau.edu.cn/PlantGSEA/analysis.php
- **MapMan**: http://mapman.gabipd.org/web/guest/mapman-download
- **ToppGene**: http://toppgene.cchmc.org/prioritization.jsp
- **gProfiler**: http://biit.cs.ut.ee/gprofiler/
- **PLAZA**: http://bioinformatics.psb.ugent.be/plaza/ (*developed at VIB*)
- **BIOMART**: http://www.biomart.org/biomart/martview
- **QuickGO**: http://www.ebi.ac.uk/ego/
- **TAIR**: http://www.arabidopsis.org/tools/bulk/index.jsp
- **BAR**: http://bar.utoronto.ca/welcome.htm
- **BioCyc**: http://biocyc.org/gene-search.shtml
  - **BioCyc Ath -> pathways**: http://biocyc.org/ARA/class-tree?object=Pathways

Meta Analysis Resources

- **InsilicoDB** [9] offers similar services by linking the data to the Broad data-mining tools **GenePattern** & **Gene-E**. Please refer to the InsilicoDB **tutorial pages (https://insilicodb.com/category/tutorials/)** for more info.

Commercial resources licensed by VIB

- **Genevestigator** not covered during this training but warmly recommended for all users who do not have their own MA data but need to find biomarkers.
- **CLC Main workbench** (http://data.bits.vib.be/pub/trainingen/CLCMain/TutorialMicroarrays.pdf) used in the optional PubMA_Exercise.6
- **Ingenuity Pathway Analysis (IPA)** is strongly advised for more advanced users/usage. You can use IPA on any Java-installed computer after asking for a personal account to **mailto:bits@vib.be** and login in **here (https://apps.ingenuity.com/ingsso/login?service=https%3A%2F%2Fanalysis.ingenuity.com%2Fpa%2Fj_spring_cas_security_check&originalUrl=https%3A%2F%2Fanalysis.ingenuity.com%2Fpa%3Futm_source%3DIngenuity%26utm_medium%3DWebsite%26utm_campaign%3DIPALoginPage)** . **Please keep in mind that IPA is only meant for human/mouse/rat data**.

Do you still need MORE?

Find more tools with **OMICtools (http://omictools.com)**[10]

References:

1. ↑

   Tanya Barrett, Ron Edgar
   Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*.
   Methods Mol. Biol.: 2006, 338;175-90
   [PubMed:16888359] ##WORLDCAT## [DOI] (P p)

   Ron Edgar, Michael Domrachev, Alex E Lash
   Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.
   Nucleic Acids Res.: 2002, 30(1);207-10
   [PubMed:11752295] ##WORLDCAT## (I p)

2. ↑ http://www.vib.be/en/training/research-training/courses/Pages/Analysis-of-public-microarray-data-sets.aspx
3. ↑ http://www.vib.be/en/training/research-training/courses/Pages/Introduction-to-Affymetrix-microarray-analysis.aspx http://www.vib.be/en/training/research-training/courses/Pages/Analysis-of-public-microarray-data-using-Genevestigator.aspx
4. ↑ https://www.bits.vib.be/index.php/training/177-microarray-bioconductor https://www.bits.vib.be/index.php/training/125-genevestigator
5. ↑ http://genepattern.org/
6. ↑ http://www.broadinstitute.org/cancer/software/GENE-E/
7. ↑ http://www.broadinstitute.org/gsea/
8. ↑ http://tagc.univ-mrs.fr/tbrowser/

   Cyrille Lepoivre, Aurélie Bergon, Fabrice Lopez, Narayanan B Perumal, Catherine Nguyen, Jean Imbert, Denis Puthier
   TranscriptomeBrowser 3.0: introducing a new compendium of molecular interactions and a new visualization tool for the study of gene regulatory networks.
   BMC Bioinformatics: 2012, 13;19
   [PubMed:22292669] ##WORLDCAT## [DOI] (I e)

   Fabrice Lopez, Julien Textoris, Aurélie Bergon, Gilles Didier, Elisabeth Remy, Samuel Granjeaud, Jean Imbert, Catherine Nguyen, Denis Puthier
   TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database.
   PLoS ONE: 2008, 3(12);e4001
   [PubMed:19104654] ##WORLDCAT## [DOI] (I p)

9. ↑ https://insilicodb.com InsilicoDB

   Jonatan Taminau, Stijn Meganck, Cosmin Lazar, David Steenhoff, Alain Coletta, Colin Molter, Robin Duque, Virginie de Schaetzen, David Y Weiss Solís, Hugues Bersini, Ann Nowé
   Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages.
   BMC Bioinformatics: 2012, 13;335
   [PubMed:23259851] ##WORLDCAT## [DOI] (I e)

   Alain Coletta, Colin Molter, Robin Duqué, David Steenhoff, Jonatan Taminau, Virginie de Schaetzen, Stijn Meganck, Cosmin Lazar, David Venet, Vincent Detours, Ann Nowé, Hugues Bersini, David Y Weiss Solís
   InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor.
   Genome Biol.: 2012, 13(11);R104
   [PubMed:23158523] ##WORLDCAT## [DOI] (I e)

   Cosmin Lazar, Stijn Meganck, Jonatan Taminau, David Steenhoff, Alain Coletta, Colin Molter, David Y Weiss-Solís, Robin Duque, Hugues Bersini, Ann Nowé
   Batch effect removal methods for microarray gene expression data integration: a survey.
   Brief. Bioinformatics: 2013, 14(4);469-90
   [PubMed:22851511] ##WORLDCAT## [DOI] (I p)

   Jonatan Taminau, David Steenhoff, Alain Coletta, Stijn Meganck, Cosmin Lazar, Virginie de Schaetzen, Robin Duque, Colin Molter, Hugues Bersini, Ann Nowé, David Y Weiss Solís
   inSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO.
   Bioinformatics: 2011, 27(22);3204-5
   [PubMed:21937664] ##WORLDCAT## [DOI] (I p)

10. ↑ http://omictools.com

    Vincent J Henry, Anita E Bandrowski, Anne-Sophie Pepin, Bruno J Gonzalez, Arnaud Desfeux
    OMICtools: an informative directory for multi-omic data analysis.
    Database (Oxford): 2014, 2014;
    [PubMed:25024350] ##WORLDCAT## [DOI] (I e)

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=Hands-on_Analysis_of_public_microarray_datasets&oldid=11837"
Categories: Training | HandsOn | PUBMA2014

- This page was last modified on 21 October 2014, at 10:39.

This page was last modified on 27 October 2014, at 08:37.
This page has been accessed 289 times.
Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

PubMA Exercise.1

From BioWareWIKI

Search **GEO** to find public datasets related to one's project

---

Contents

- 1 Introduction
- 2 Find GEO datasets relevant for your Biological question
- 3 Get information about GSE6943 used in this session
- 4 download exercise files

Introduction

<<The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the studies and gene expression patterns stored in GEO. For more information about various aspects of GEO, please see our documentation (http://www.ncbi.nlm.nih.gov/geo/info/) listings and publications (http://www.ncbi.nlm.nih.gov/pmc/3013736,2686538,2270403,1669752,1619900,1619899,539976,99122)>> [1].

If you are new to GEO have a look to their **handout (http://www.ncbi.nlm.nih.gov/geo/info/GEOHandoutFinal.pdf)** first ![2]

Find GEO datasets relevant for your Biological question

You may use the NCBI search page in a very basic way by entering your gene of interest to look for related knock-out experiments or by searching for a compound or disease name that is relevant to your research. Note that this will sometimes find too many datasets or miss the true one you dream of.

Other, better ways, of finding if GEO data does exist

- The current NCBI search page (and its advanced counterpart) also allows restricting your queries in smart ways and reach the goal of finding the best suited data in the repository. A related How-To page be found at **Find_GEO_datasets**. Please read this page and discover the advantages of adding filters to your queries.
- For a good resource about how to build top-notch queries in the **GEO advanced search page (http://www.ncbi.nlm.nih.gov/gds/advanced)** look at the NCBI tutorial [3] with examples of good syntax to recycle and copy [4]. As example, search experiments in rat with more than 100 samples with '(100:500[Number of Samples]) AND rat[Organism]'.

Get information about GSE6943 used in this session

We will stick to the simple NCBI-GEO search here and look for one particular experiment (GSE6943) used by CLC to build their tutorial. This work was published by 'van Lunteren E, Spiegler S, Moyer M' [5]. The sequencing was performed on tumor cultures from 4 patients at 2 time points over 3 conditions (DPN, OHT and control). One control sample was omitted by the paper authors due to low quality[6]. Full details about this dataset can be found at **http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=+GSE6943** and is reproduced in the next figure. A lot of important information can be found on this page including the chip used for the experiment, the number of replicates, as well as metadata information about the experimental setup.

Series GSE6943 Query DataSets for GSE6943

Status	Public on Jan 24, 2008
Title	Normal Heart vs Normal Diaphragm
Organism	Rattus norvegicus
Experiment type	Expression profiling by array
Summary	Comparison of gene expression of heart (left vent) and diaphragm of normal Sprague Dawley rats, young adult Keywords: Cell type comparison
Overall design	6 diaphragm samples 6 heart samples
Contributor(s)	van Lunteren E, Spiegler S, Moyer M
Citation(s)	van Lunteren E, Spiegler S, Moyer M. Contrast between cardiac left ventricle and diaphragm muscle in expression of genes involved in carbohydrate and lipid metabolism. *Respir Physiol Neurobiol* 2008 Mar 20;161(1):41-53. PMID: 18207466
Submission date	Feb 02, 2007
Last update date	Jun 21, 2012
Contact name	Erik -- van Lunteren
E-mail	exv4@cwru.edu
Phone	216-791-3800
Organization name	Cleveland VA Medical Center
Department	Pulmonary 111J(W)
Street address	10701 East Boulevard
City	Cleveland
State/province	OH
ZIP/Postal code	44106
Country	USA
Platforms (1)	GPL341 [RAE230A] Affymetrix Rat Expression 230A Array
Samples (12) More...	GSM160089 Diaphragm 1 GSM160090 Diaphragm 2 GSM160091 Diaphragm 3

Relations
BioProject PRJNA98125 -> clusters & heatmaps

Analyze with GEO2R -> compute DE

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE6943_RAW.tar	16.8 Mb	(http)(custom)	TAR (of CEL)
Raw data provided as supplementary file			
Processed data included within Sample table			

Note the two red boxes that highlight links to two additional resources at GEO

- The first link directs the user to the **PRJNA98125 (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA98125)** BioProject page related to this submission and including extra links to the corresponding DataSet. Note that only those submissions pre-procesed by the GEO team are linked to a DataSet (read **http://www.ncbi.nlm.nih.gov/geo/info/datasets.html**). DataSets form the basis of GEO's advanced data display and analysis tools, including gene expression profile charts and clusters detailed in **PubMA_Exercise.3**.
- The second link **Analyze with GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE6943)** opens a new window with the **GEO2R** submission form further detailed in the next **PubMA_Exercise.2**.

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑ http://www.ncbi.nlm.nih.gov/geo/info/faq.html#What
2. ↑ http://www.ncbi.nlm.nih.gov/geo/info/GEOHandoutFinal.pdf
3. ↑ http://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html
4. ↑ http://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html#fields
5. ↑

   Erik van Lunteren, Sarah Spiegler, Michelle Moyer
   Contrast between cardiac left ventricle and diaphragm muscle in expression of genes involved in carbohydrate and lipid metabolism.
   Respir Physiol Neurobiol: 2008, 161(1);41-53
   [PubMed:18207466] ##WORLDCAT## [DOI] (P p)

   http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=+GSE6943

6. ↑ http://www.bioconductor.org/packages/release/data/experiment/html/parathyroidSE.html

---

[Main_Page | Hands-on Analysis of public microarray datasets | PubMA_Exercise.2]

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=PubMA_Exercise.1&oldid=11745"
Category: PUBMA2014

- This page was last modified on 16 October 2014, at 08:49.
- This page has been accessed 133 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

PubMA Exercise.2

From BioWareWIKI

Compute differential analysis using GEO2R within the NCBI web-portal

---

Contents

- 1 Analyze public GEO data on the NCBI portal
  - 1.1 GEO2R step-by-step walk-through for GSE6943
    - 1.1.1 The GEO2R interface
    - 1.1.2 GEO2R sample definition
    - 1.1.3 Visualize the distribution of log-transformed expression values
    - 1.1.4 Search for the top 250 differentially expressed transcripts
    - 1.1.5 Saving the Rscript for further use in RStudio
- 2 download exercise files

Analyze public GEO data on the NCBI portal

The GEO portal links to several web-tools allowing data analysis without the need to install anything on your computer. Although these tools will not compete with sophisticated R/Bioconductor methods, they remain very attractive as they do not require prior knowledge in MA data analysis and are very fast, leading the users to tabular results and pictures that can be fed to other tools or used as is in scientific reports. We proceed here with **GEO2R** which allows finding differentially expressed genes by comparing sample groups within one GEO submission. Full instructions (https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html)[1] - (Tutorial video

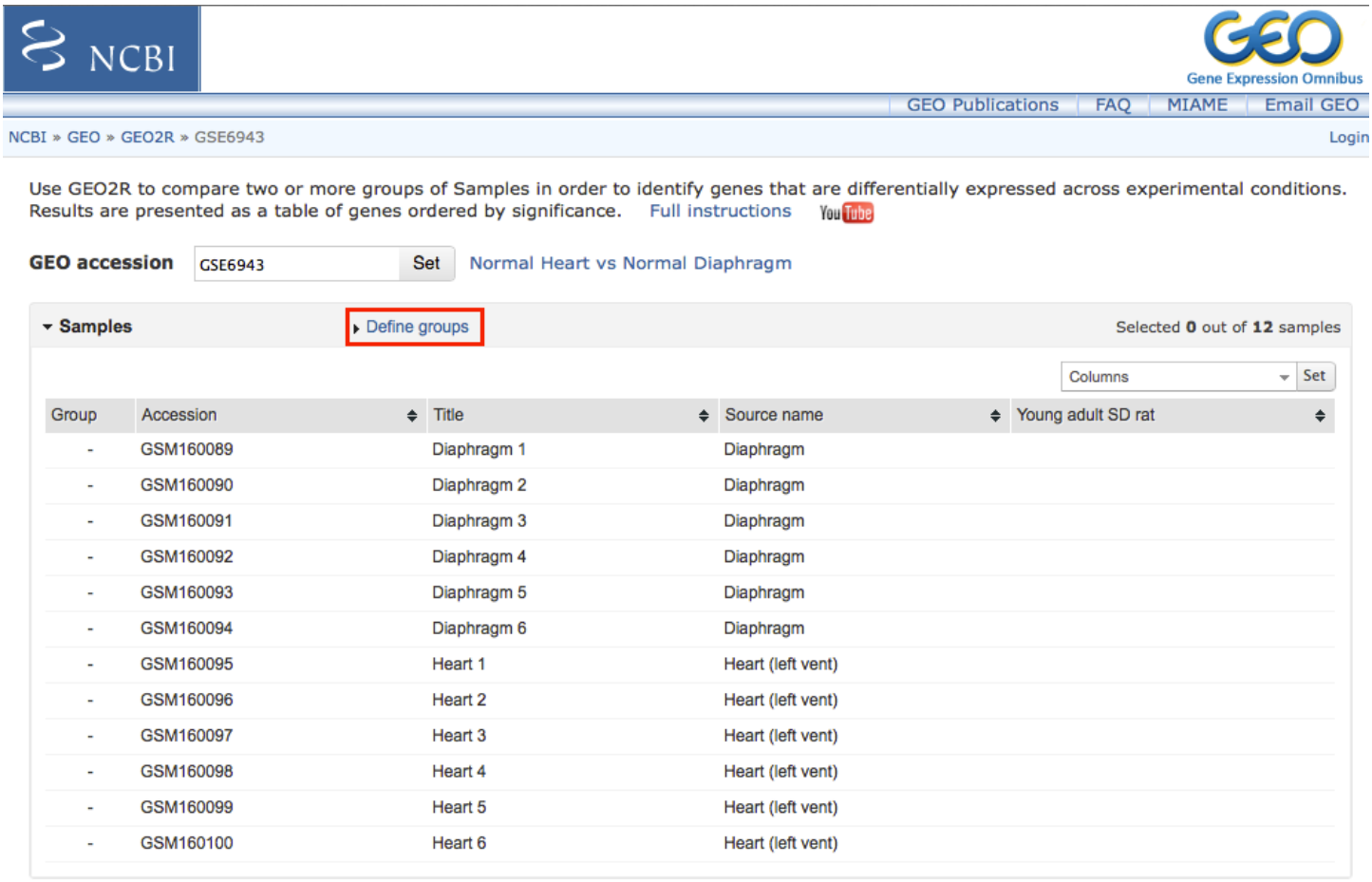
)

GEO2R step-by-step walk-through for GSE6943

Although former HowTo page Analyze_GEO_data_with_GEO2R is already present on the BITS Wiki, we repeat the analysis here with the same dataset used in the CLC main workbench to be able to compare results of free and commercial solutions. This work was published by 'van Lunteren E, Spiegler S, Moyer M' [2]. Full details about this dataset can be found on the **http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=+GSE6943** GEO page.

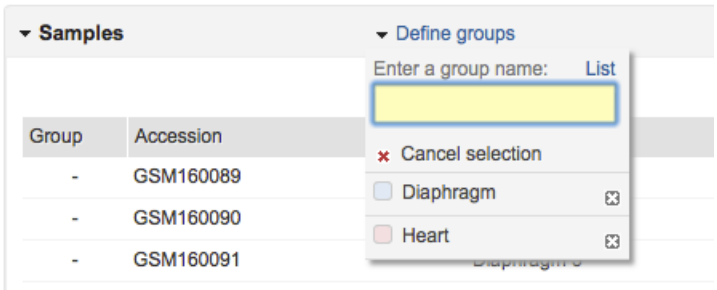
The GEO2R interface

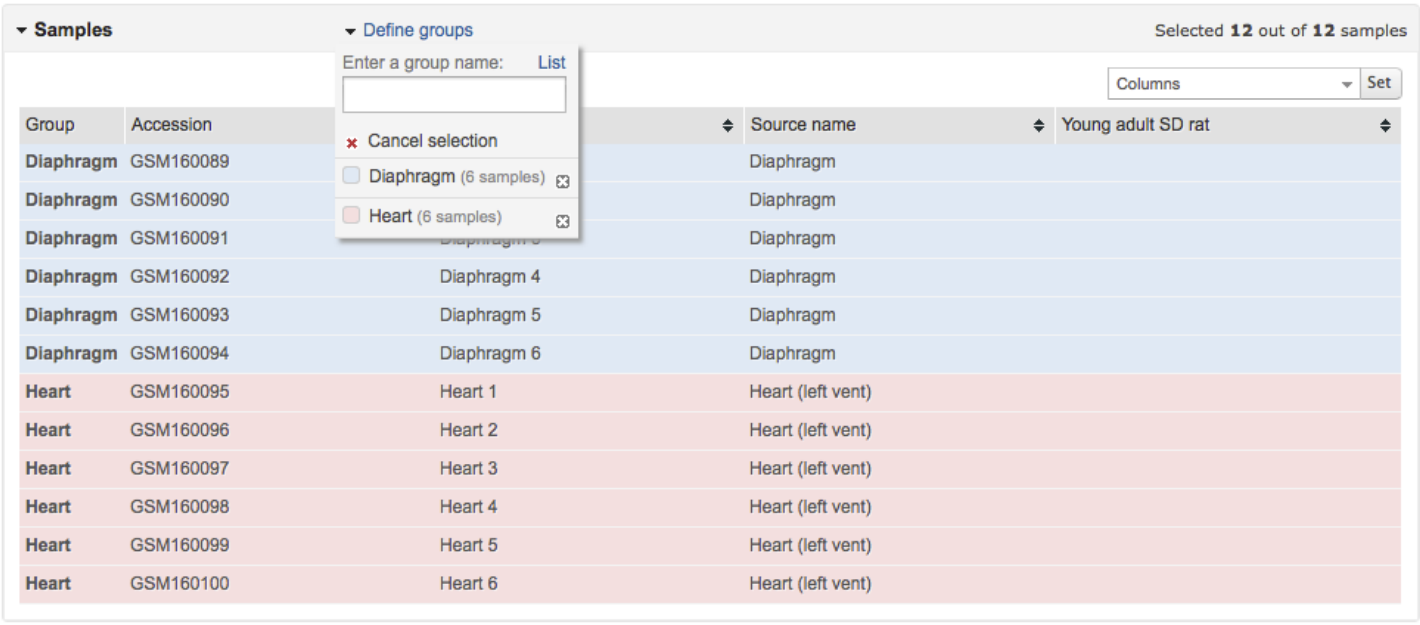
The initial window shows several TABs that will be reviewed in the remaining of this tutorial.


NCBI
Gene Expression Omnibus
GEO Publications
FAQ
MIAME
Email GEO
NCBI » GEO » GEO2R » GSE6943
Login
Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions.
Results are presented as a table of genes ordered by significance.
Full instructions
GEO accession
GSE6943
Set
Normal Heart vs Normal Diaphragm
Samples
Define groups
Selected 0 out of 12 samples
Columns
Set
Group
Accession
Title
Source name
Young adult SD rat
- GSM160089 Diaphragm 1 Diaphragm
- GSM160090 Diaphragm 2 Diaphragm
- GSM160091 Diaphragm 3 Diaphragm
- GSM160092 Diaphragm 4 Diaphragm
- GSM160093 Diaphragm 5 Diaphragm
- GSM160094 Diaphragm 6 Diaphragm
- GSM160095 Heart 1 Heart (left vent)
- GSM160096 Heart 2 Heart (left vent)
- GSM160097 Heart 3 Heart (left vent)
- GSM160098 Heart 4 Heart (left vent)
- GSM160099 Heart 5 Heart (left vent)
- GSM160100 Heart 6 Heart (left vent)

GEO2R sample definition

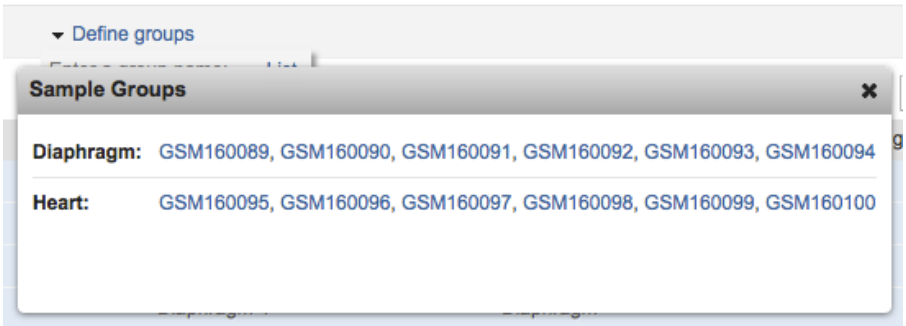
The first step in the GEO2R analysis is performed by cliscking on **Define groups** to setup sample groups based on available samples and label them. These groups will be used to define contrasts and compute pairwise differential expression analyses. Two groups are created with names 'diaphragm' and 'heart' then samples are labeled using the mouse.





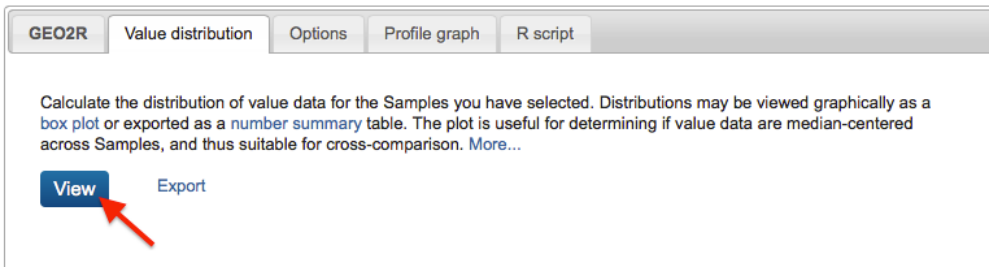
The order in which you assign the groups is important. First define the control group (it will be colored in blue), then define the treated group (it will be colored in pink). The order is important for calculating log fold changes later in the analysis. If you reverse the order: genes that are upregulated according to the publication that supports the data will be downregulated in your results and vice versa

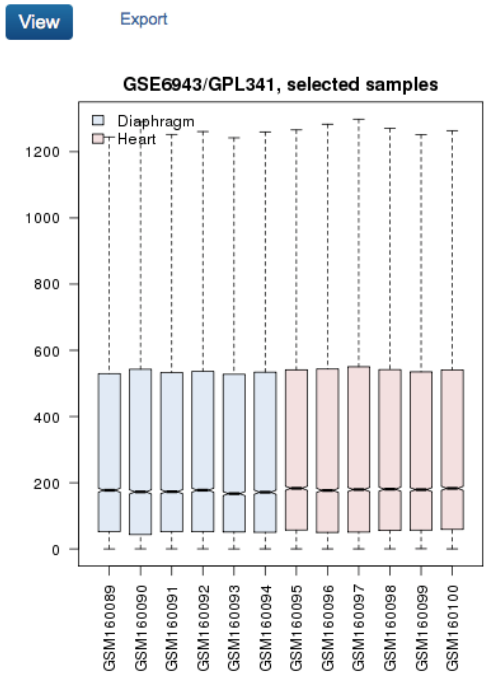
The list of samples in each group can be reviewed by clicking on **List** in the group definition popup window.


Define groups
Sample Groups
Diaphragm: GSM160089, GSM160090, GSM160091, GSM160092, GSM160093, GSM160094
Heart: GSM160095, GSM160096, GSM160097, GSM160098, GSM160099, GSM160100

Visualize the distribution of log-transformed expression values

Before proceeding with DE analysis, it is very important to first control for sample value distribution homogeneity in the 'Value distribution' TAB.


GEO2R
Value distribution
Options
Profile graph
R script
Calculate the distribution of value data for the Samples you have selected. Distributions may be viewed graphically as a box plot or exported as a number summary table. The plot is useful for determining if value data are median-centered across Samples, and thus suitable for cross-comparison. More...
View
Export



This plot represents the distribution of the data in all samples.

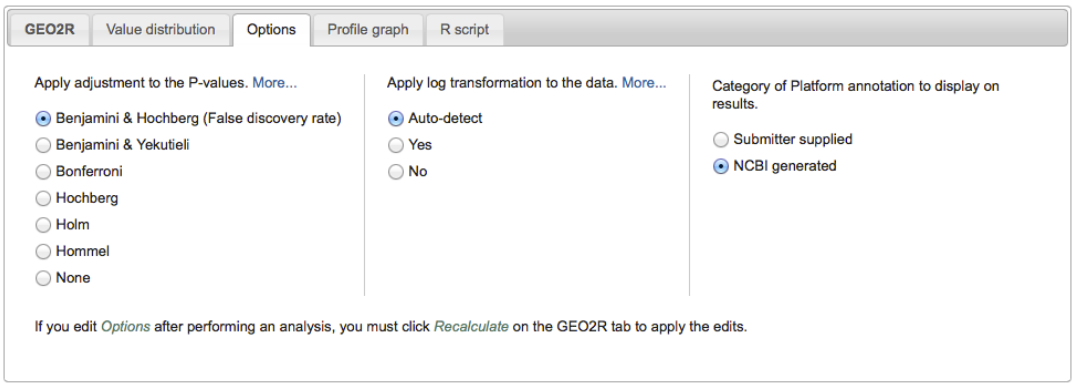
Since the data is supposed to be normalized you expect comparable boxes for all samples. When box plots show large divergence, it might point to the fact that the data in the Series Matrix file was not yet normalized. Unfortunately you cannot perform normalization in GEO2R. If the boxes are very different, then it is not possible to compare the samples.

Search for the top 250 differentially expressed transcripts

Since the boxplots show that the data has been normalized, we can now proceed with finding DE genes (top-250 being a good proxy for downstream analysis) between the two groups.

*Options can be set in the **Options** tab to handle log transformation and multiple testing correction to be applied to the data.*

The default Options are shown below and are the best choice for most data sets



When satisfied, go to the **GEO2R** tab and click the **Top250** button to run a limma analysis for identifying DE genes.

*When more than two grous are defined, GEO2R selects pairwise contrasts in a triangular/circular way (depending on the number of groups). These contrasts are labelled with arbitrary names (G0, G1, ... Gn) and do not always reflect the user expectation but there is unfortunately little to be done in GEO2R to control this choice; BUT more can be done when post-processing the code in **RStudio** as will be shown in the dedicated tutorial*

GEO2R | Value distribution | Options | Profile graph | R script

▸ Quick start

Log-transformation has been applied to the data. You can change this in the Options tab.

Recalculate if you changed any options. Save all results Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
▸ 1388876_at	7.67e-12	7.81e-16	59.1	25.3	7.14		
▸ 1374248_at	7.67e-12	9.78e-16	-57.9	25.1	-11.48	Mybpc1	myosin binding p…
▸ 1374622_at	7.67e-12	1.44e-15	-56	24.9	-4.69		
▸ 1370033_at	3.47e-11	8.71e-15	-48	23.6	-7.79	Myl1	myosin, light chai…
▸ 1371339_at	4.62e-11	1.45e-14	-46	23.2	-8.25	Tnni1	troponin I type 1 …
▸ 1373697_at	7.19e-11	2.71e-14	-43.6	22.7	-7.13	Mybpc2	myosin binding p…
▸ 1398306_at	9.89e-11	4.35e-14	-41.8	22.3	-7.06	Ampd1	adenosine mono…
▸ 1374672_at	1.14e-10	5.70e-14	40.8	22.1	8.82	Tnni3k	TNNI3 interactin…
▸ 1367962_at	1.44e-10	8.13e-14	-39.6	21.8	-11.53	Actn3	actinin alpha 3
▸ 1367964_at	1.44e-10	9.53e-14	-39.1	21.7	-8.57	Tnni2	troponin I type 2 …
▸ 1376227_at	1.44e-10	1.06e-13	-38.7	21.6	-6.48	Myoz1	myozenin 1
▸ 1387065_at	1.44e-10	1.09e-13	-38.6	21.6	-5.85	Plcd4	phospholipase C,…
▸ 1384202_at	2.03e-10	1.67e-13	37.2	21.2	6.68	Tesc	tescalcin
▸ 1386931_at	2.03e-10	1.78e-13	37	21.2	6.71	Tnni3	troponin I type 3 …
▸ 1384178_at	2.56e-10	2.42e-13	36.1	20.9	4.38	Lrrc4b	leucine rich repe…
▸ 1371298_at	2.56e-10	2.57e-13	-35.9	20.8	-6.34	H19	H19, imprinted m…
▸ 1367604_at	3.12e-10	3.33e-13	35.1	20.6	3.56	Crip2	cysteine-rich prot…
▸ 1367896_at	3.70e-10	4.18e-13	-34.4	20.4	-8.49	Car3	carbonic anhydra…
▸ 1375739_at	3.71e-10	4.43e-13	34.2	20.4	3.46	Ehd4	EH-domain conta…
▸ 1389532_at	4.67e-10	5.98e-13	33.4	20.1	5.07	Nebl	nebulette

The limma analysis results in a list of the 250 transcripts with the lowest p-values (ranked by increasing p-value).
The results table contains the following columns:

- adj.P.Val: p-value after correction for multiple testing.
  This column is the **statistic you should use for interpreting the results**. Genes with the smallest adjusted p-values will be the most reliable. Selecting all transcripts with adjusted p-values < 0.05 is equivalent to setting the False Discovery Rate (FDR) to 0,05 allowing that 5% of the selected DE genes are false positives.
  As you can see GEO2R always shows the 250 genes with the lowest p-values, regardless of the significance of their p-values. Sometimes 250 is not enough and you still miss DE genes (as is the case in this example), sometimes 250 is way too much and only a small fraction of these 250 genes is really DE. So always check the adjusted p-values to decide how many genes of these 250 you are going to use for further analysis.
- P.Value: raw p-value before multiple testing correction
- t: t-statistic of the shrunken t-test
- B: B-statistic or log-odds that the gene is differentially expressed
- logFC: Log2-fold change between the two experimental conditions

This table contains links through which detailed expression information can be retrieved for interesting genes (not further detailed here).

Clicking on 'Save all results' will open a new window with the full table that can be saved to disk as a tab-separated text file using the browser File Save option

GEO2R - GEO - NCBI | www.ncbi.nlm.nih.gov/geo/geo2r/backend/geo2r.cgi?ctg_time=14078...

```
"ID"    "adj.P.Val"     "P.Value"       "t"     "B"     "logFC" "Gene.symbol"   "Gene.title"
"1388876_at"    "7.67e-12"      "7.81e-16"      "5.91e+01"      "25.28866"      "7.14"  ""      ""
"1374248_at"    "7.67e-12"      "9.78e-16"      "-5.79e+01"     "25.14116"      "-1.15e+01"     "Mybpc1"        "myosin
binding protein C, slow type"
"1374622_at"    "7.67e-12"      "1.44e-15"      "-5.60e+01"     "24.87971"      "-4.69" ""      ""
"1370033_at"    "3.47e-11"      "8.71e-15"      "-4.80e+01"     "23.59223"      "-7.79" "Myl1"  "myosin, light chain 1"
"1371339_at"    "4.62e-11"      "1.45e-14"      "-4.60e+01"     "23.20335"      "-8.25" "Tnni1" "troponin I type 1 (skeletal,
slow)"
"1373697_at"    "7.19e-11"      "2.71e-14"      "-4.36e+01"     "22.71378"      "-7.13" "Mybpc2"        "myosin binding
protein C, fast-type"
"1398306_at"    "9.89e-11"      "4.35e-14"      "-4.18e+01"     "22.33291"      "-7.06" "Ampd1" "adenosine monophosphate
deaminase 1"
"1374672_at"    "1.14e-10"      "5.70e-14"      "4.08e+01"      "22.11085"      "8.82"  "Tnni3k"        "TNNI3 interacting
kinase"
"1367962_at"    "1.44e-10"      "8.13e-14"      "-3.96e+01"     "21.81731"      "-1.15e+01"     "Actn3" "actinin alpha 3"
"1367964_at"    "1.44e-10"      "9.53e-14"      "-3.91e+01"     "21.6842"       "-8.57" "Tnni2" "troponin I type 2 (skeletal,
fast)"
"1376227_at"    "1.44e-10"      "1.06e-13"      "-3.87e+01"     "21.59864"      "-6.48" "Myoz1" "myozenin 1"
"1387065_at"    "1.44e-10"      "1.09e-13"      "-3.86e+01"     "21.57382"      "-5.85" "Plcd4" "phospholipase C, delta 4"
"1384202_at"    "2.03e-10"      "1.67e-13"      "3.72e+01"      "21.20727"      "6.68"  "Tesc"  "tescalcin"
"1386931_at"    "2.03e-10"      "1.78e-13"      "3.70e+01"      "21.15193"      "6.71"  "Tnni3" "troponin I type 3 (cardiac)"
"1384178_at"    "2.56e-10"      "2.42e-13"      "3.61e+01"      "20.88785"      "4.38"  "Lrrc4b"        "leucine rich repeat
containing 4B"
"1371298_at"    "2.56e-10"      "2.57e-13"      "-3.59e+01"     "20.83498"      "-6.34" "H19"   "H19, imprinted maternally
expressed transcript (non-protein coding)"
"1367604_at"    "3.12e-10"      "3.33e-13"      "3.51e+01"      "20.6094"       "3.56"  "Crip2" "cysteine-rich protein 2"
"1367896_at"    "3.70e-10"      "4.18e-13"      "-3.44e+01"     "20.40929"      "-8.49" "Car3"  "carbonic anhydrase 3"
"1375739_at"    "3.71e-10"      "4.43e-13"      "3.42e+01"      "20.35831"      "3.46"  "Ehd4"  "EH-domain containing 4"
"1389532_at"    "4.67e-10"      "5.98e-13"      "3.34e+01"      "20.09179"      "5.07"  "Nebl"  "nebulette"
"1370198_at"    "4.67e-10"      "6.44e-13"      "-3.31e+01"     "20.02504"      "-5.15" "Trdn"  "triadin"
"1370157_at"    "4.67e-10"      "6.45e-13"      "3.31e+01"      "20.02293"      "6.31"  "Pln"   "phospholamban"
"1386873_at"    "5.30e-10"      "7.66e-13"      "-3.26e+01"     "19.86935"      "-7.37" "Tnni1" "troponin I type 1 (skeletal,
slow)"
```

*If you wish to upload this table to Ingenuity Pathway Analysis (**IPA**), you may consider opening it first in Microsoft Excel and save it back as a '.xls' file. This will remove the double quotes around fields and allow better recognition of your data by IPA*

Saving the Rscript for further use in RStudio

This is the last step of this tutorial and the first step of the follow-up page **PubMA_Exercise.2b** where we will produce an R script to perform the GEO2R analysis on our own computer and prepare for more advanced microarray analyses.

GEO2R | Value distribution | Options | Profile graph | R script

```
# Version info: R 2.14.1, Biobase 2.15.3, GEOquery 2.23.2, limma 3.10.1
# R scripts generated  Tue Aug 12 05:30:54 EDT 2014

################################################################
#   Differential expression analysis with limma
library(Biobase)
library(GEOquery)
library(limma)

# load series and platform data from GEO

gset <- getGEO("GSE6943", GSEMatrix =TRUE)
if (length(gset) > 1) idx <- grep("GPL341", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group names for all samples
sml <- c("G0","G0","G0","G0","G0","G0","G1","G1","G1","G1","G1","G1");

# log2 transform
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
          (qx[6]-qx[1] > 50 && qx[2] > 0) ||
          (qx[2] > 0 && qx[2] < 1 && qx[4] > 1 && qx[4] < 2)
if (LogC) { ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex) }
```

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑ https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html

2. ↑

Erik van Lunteren, Sarah Spiegler, Michelle Moyer
Contrast between cardiac left ventricle and diaphragm muscle in expression of genes involved in carbohydrate and lipid metabolism.
Respir Physiol Neurobiol: 2008, 161(1);41-53
[PubMed:18207466] ##WORLDCAT## [DOI] (P p)

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=+GSE6943

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=PubMA_Exercise.2&oldid=11740"
Category: PUBMA2014

- This page was last modified on 15 October 2014, at 13:58.
- This page has been accessed 94 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

PubMA Exercise.2b

From BioWareWIKI

Follow-up analysis demo in RStudio

Contents

- 1 Introduction
- 2 installing on a non-BITS laptop
- 3 Extend the GEO2R analysis in R with RStudio
  - 3.1 adapt the GEO2R script in RStudio
    - 3.1.1 Original GEO2R code
    - 3.1.2 Improved code
- 4 download exercise files

Introduction

The former exercise produced a full table with differential expression between Heart and Diaphragm samples that can be used directly with Functional enrichment tools or IPA. The R-script used by GEO2R is quite standard and basic and only provides BoxPlots for signal distribution. This is a minimum when doing MA data analysis and users often appreciate to have some more QC done on the data to control for biases or inconsistencies associated with artifacts or with variability in the experiment.

Expert analyst makes use of the [R]-Bioconductor toolbox to further analyze their data. Some examples of standard R functions are shown below to show you the power of this programing language. The following exercise is provided as an appetizer and is far not exhaustive, you will find many more methods and tools by exploring the CRAN documentation (http://cran.r-project.org[1])

installing on a non-BITS laptop

required for non-BITS laptops *[Collapse]*

```
# If running a unix OS computer please use yum to install
libxml2-devel libcurl-devel #required to build dependencies
pandoc # required for markdown
texlive-collection-latexrecommended.noarch texlive-latex.noarch # also required for markdown

# install minimal bioconductor
source("http://bioconductor.org/biocLite.R")
biocLite()

# Add required packages for today's session (will add RCurl, XML, ... dependencies)
## biocLite(c("Biobase", "GEOquery", "limma",  "affy"))

# some other packages are not required here but are required for RobiNA in other exercises
# these required packages were identified by parsing all R scripts present in the windows build of Robina
## biocLite (c("affy", "affyPLM",  "RankProd", "limma", "gcrma", "statmod", "marray", "plier", "makecdfen
    "edgeR", "DESeq", "EDASeq"))
```

Extend the GEO2R analysis in R with RStudio

RStudio is the current best graphical environment to program i the [R] language and offers many facilities to learn and develop your R skills. The program is available free of charge from **http://www.rstudio.com**[2].

adapt the GEO2R script in RStudio

Original GEO2R code

The starting script is first reproduced here

initial GEO2R code *[Collapse]*

```
# Version info: R 2.14.1, Biobase 2.15.3, GEOquery 2.23.2, limma 3.10.1
# R scripts generated  Tue Aug 12 05:30:54 EDT 2014

################################################################
#   Differential expression analysis with limma
library(Biobase)
library(GEOquery)
library(limma)

# load series and platform data from GEO
## gset <- getGEO("GSE6943", GSEMatrix =TRUE)
gset <- getGEO("GSE6943", GSEMatrix =TRUE)
if (length(gset) > 1) idx <- grep("GPL341", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group names for all samples
sml <- c("G0","G0","G0","G0","G0","G0","G1","G1","G1","G1","G1","G1");

# log2 transform
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
  (qx[6]-qx[1] > 50 && qx[2] > 0) ||
  (qx[2] > 0 && qx[2] < 1 && qx[4] > 1 && qx[4] < 2)
if (LogC) { ex[which(ex <= 0)] <- NaN
            exprs(gset) <- log2(ex) }

# set up the data and proceed with analysis
fl <- as.factor(sml)
gset$description <- fl
design <- model.matrix(~ description + 0, gset)
colnames(design) <- levels(fl)
fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(G1-G0, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)

# load NCBI platform annotation
gpl <- annotation(gset)
## platf <- getGEO(gpl, AnnotGPL=TRUE)
platf <- getGEO(gpl, AnnotGPL=TRUE)
ncbifd <- data.frame(attr(dataTable(platf), "table"))

# replace original platform annotation
tT <- tT[setdiff(colnames(tT), setdiff(fvarLabels(gset), "ID"))]
tT <- merge(tT, ncbifd, by="ID")
tT <- tT[order(tT$P.Value), ]  # restore correct order

tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene.title"))
write.table(tT, file=stdout(), row.names=F, sep="\t")

################################################################
#   Boxplot for selected GEO samples
library(Biobase)
library(GEOquery)

# load series and platform data from GEO

## gset <- getGEO("GSE6943", GSEMatrix=TRUE)
gset <- getGEO("GSE6943", GSEMatrix=TRUE)
if (length(gset) > 1) idx <- grep("GPL341", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]
```

```
# group names for all samples in a series
sml <- c("G0","G0","G0","G0","G0","G0","G1","G1","G1","G1","G1","G1")

# order samples by group
ex <- exprs(gset)[ , order(sml)]
sml <- sml[order(sml)]
fl <- as.factor(sml)
labels <- c("Diaphragm","Heart")

# set parameters and draw the plot
palette(c("#dfeaf4","#f4dfdf", "#AABBCC"))
dev.new(width=4+dim(gset)[[2]]/5, height=6)
par(mar=c(2+round(max(nchar(sampleNames(gset)))/2),4,2,1))
title <- paste ("GSE6943", '/', annotation(gset), " selected samples", sep ='')
boxplot(ex, boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=fl)
legend("topleft", labels, fill=palette(), bty="n")
```

Improved code

The next script was adapted to keep local files and to save results to file instead of sanding them to a new browser window. Edits are shown where original lines are commented out by '##'

- 'destdir = base' saves the downloads to the local folder instead of to the temp directory
- 'number=nrow(fit2)' generates the full table instead of the top250

We now show the edited script where lines starting with ## are modified in the next line(s)

```
## added configuration
base <- "~/PUBMA2014/ex2b-files", sep="")
setwd(base)

# Version info: R 2.14.1, Biobase 2.15.3, GEOquery 2.23.2, limma 3.10.1
# R scripts generated  Tue Aug 12 05:30:54 EDT 2014

################################################################
#   Differential expression analysis with limma
library(Biobase)
library(GEOquery)
library(limma)

# load series and platform data from GEO
## gset <- getGEO("GSE6943", GSEMatrix =TRUE)
gset <- getGEO("GSE6943", GSEMatrix =TRUE, destdir = base)
if (length(gset) > 1) idx <- grep("GPL341", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group names for all samples
sml <- c("G0","G0","G0","G0","G0","G0","G1","G1","G1","G1","G1","G1");

# log2 transform
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
          (qx[6]-qx[1] > 50 && qx[2] > 0) ||
          (qx[2] > 0 && qx[2] < 1 && qx[4] > 1 && qx[4] < 2)
if (LogC) { ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex) }

# set up the data and proceed with analysis
fl <- as.factor(sml)
gset$description <- fl
design <- model.matrix(~ description + 0, gset)
colnames(design) <- levels(fl)
fit <- lmFit(gset, design)
cont.matrix <- makeContrasts(G1-G0, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, 0.01)

## tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
tT <-  topTable(fit2, adjust="fdr", sort.by="B", number=nrow(fit2))

# load NCBI platform annotation
gpl <- annotation(gset)
## platf <- getGEO(gpl, AnnotGPL=TRUE)
```

```
platf <- getGEO(gpl, AnnotGPL=TRUE, destdir = base)
ncbifd <- data.frame(attr(dataTable(platf), "table"))

# replace original platform annotation
tT <- tT[setdiff(colnames(tT), setdiff(fvarLabels(gset), "ID"))]
tT <- merge(tT, ncbifd, by="ID")
tT <- tT[order(tT$P.Value), ]  # restore correct order

## tT <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene.title"))
tT.final <- subset(tT, select=c("ID","adj.P.Val","P.Value","t","B","logFC","Gene.symbol","Gene.title"))
## write.table(tT.final, file=stdout(), row.names=F, sep="\t")
write.table(tT.final, file="GSE6943_DE.txt", row.names=F, sep="\t", quote=FALSE)
###############################################################
#   Boxplot for selected GEO samples
library(Biobase)
library(GEOquery)

# load series and platform data from GEO

## gset <- getGEO("GSE6943", GSEMatrix=TRUE)
gset <- getGEO("GSE6943", GSEMatrix=TRUE, destdir = base)
if (length(gset) > 1) idx <- grep("GPL341", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# group names for all samples in a series
sml <- c("G0","G0","G0","G0","G0","G0","G1","G1","G1","G1","G1","G1")

# order samples by group
ex <- exprs(gset)[ , order(sml)]
sml <- sml[order(sml)]
fl <- as.factor(sml)
labels <- c("Diaphragm","Heart")

# set parameters and draw the plot
## save to file
filename <- paste(base, "GSE6943-boxplot.pdf", sep="/")
pdf(file = filename, bg = "white")
palette(c("#dfeaf4", "#f4dfdf", "#AABBCC"))
## dev.new(width=4+dim(gset)[[2]]/5, height=6)
par(mar=c(2+round(max(nchar(sampleNames(gset)))/2),4,2,1))
title <- paste ("GSE6943", '/', annotation(gset), " sample signal distribution", sep ='')
boxplot(ex, boxwex=0.6, notch=T, main=title, outline=FALSE, las=2, col=fl)
legend("topleft", labels, fill=palette(), bty="n")
dev.off()
```

```
## save R workspace for reuse
outfile <- paste(base, "Workspace.RData", sep="/")
save.image(file = outfile)
sessionInfo()
```

```
R version 3.1.1 (2014-07-10)
Platform: x86_64-apple-darwin10.8.0 (64-bit)

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] GEOquery_2.31.1     Biobase_2.25.0      BiocGenerics_0.11.4 limma_3.21.12

loaded via a namespace (and not attached):
[1] RCurl_1.95-4.3 tools_3.1.1    XML_3.98-1.1
```

If you want to download the raw data (CEL files) instead of the normalized data (Series Matrix file), you can adjust the script as follows:

```
# Version info: R 2.14.1, Biobase 2.15.3, GEOquery 2.23.2, limma 3.10.1
# R scripts generated  Mon Oct 13 08:43:55 EDT 2014

###############################################################
#   Differential expression analysis with limma
library(Biobase)
library(GEOquery)
library(affy)
```

```
#load CEL-files from GEO
getGEOSuppFiles("GSE6943")
untar("GSE6943/GSE6943_RAW.tar", exdir="data")
cels <- list.files("data/", pattern = "[gz]")
sapply(paste("data", cels, sep="/"), gunzip)
cels

#path to the folder in which R saved the CEL-files
#one of the files, GSM160097, has been corrupted, you have to remove it from the folder
celpath <- "C:/Users/Janick/Documents/data/"
fns <- list.celfiles(path=celpath,full.names=TRUE)
fns
cat("Reading files:\n",paste(fns,collapse="\n"),"\n")
celfiles <- ReadAffy(celfile.path=celpath)
```

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑ http://cran.r-project.org
2. ↑ http://www.rstudio.com

[Main_Page | PubMA_Exercise.1 | PubMA_Exercise.2 |
| PubMA_Exercise.3]

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=PubMA_Exercise.2b&oldid=11792"
Category: PUBMA2014

- This page was last modified on 20 October 2014, at 09:00.
- This page has been accessed 54 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

PubMA Exercise.3

From BioWareWIKI

Clustering using the **GEO Dataset browser** (only for data with attached GDS ID)

Contents

- 1 Introduction
- 2 The GEO Dataset browser
  - 2.1 Start the tool
  - 2.2 Select GEO Dataset Browser data
  - 2.3 Download the full data table
- 3 Walking through the tools
  - 3.1 Find Genes
  - 3.2 Compare two sets of samples
    - 3.2.1 Two tails comparison
      - 3.2.1.1 Profile data
      - 3.2.1.2 Profile pathways
    - 3.2.2 Rank mean difference "4x & either"
      - 3.2.2.1 Search pathways enriched in the obtained subset
  - 3.3 Cluster heatmaps
    - 3.3.1 Hierarchical clustering
    - 3.3.2 KMean clustering
  - 3.4 Experiment design and value distribution
- 4 download exercise files

Introduction

The GEO Dataset Browser [1] takes care of providing users clustered data and heatmaps derived from manual curation of some of the GEO datasets selected by the NCBI team. This processing is otherwise laborious and requires [R] skills that not every scientist can build. More info about this toolset can be found on the NCBI **help page (http://www.ncbi.nlm.nih.gov/geo/info/datasets.html)**[2]

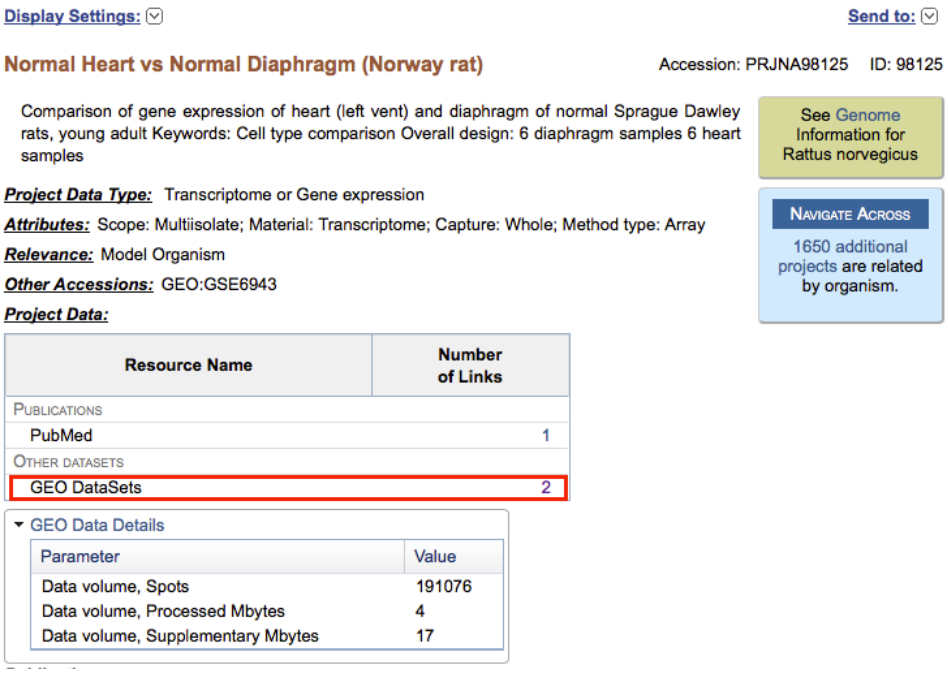
The web interface allows finding co-expressed genes in bi-clusters that have been shown to often belong to common signaling pathways or result from transcriptional co-regulation by common key regulators (TFs or pathways).

The GEO Dataset browser

The remaining of this page shows key view obtained by navigating in the **GEO Dataset browser** Web interface

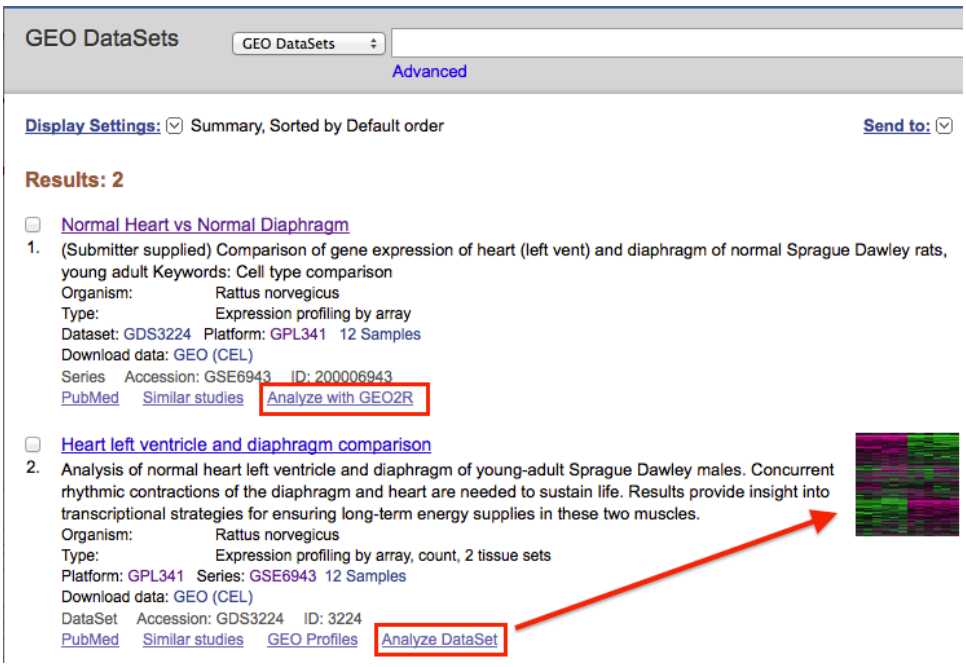
Start the tool

Start by locating the GDB link in the dataset information page found on the **http://www.ncbi.nlm.nih.gov/bioproject/PRJNA98125** GEO BioProject page or search for the GDSIDF if you know it.

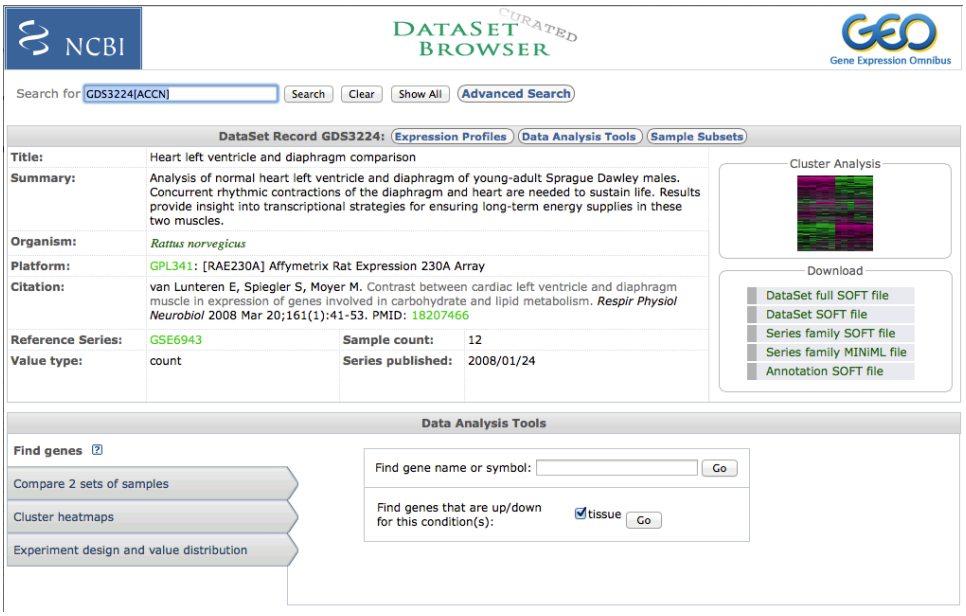

Display Settings:
Send to:
Normal Heart vs Normal Diaphragm (Norway rat)
Accession: PRJNA98125 ID: 98125
Comparison of gene expression of heart (left vent) and diaphragm of normal Sprague Dawley rats, young adult Keywords: Cell type comparison Overall design: 6 diaphragm samples 6 heart samples
See Genome Information for Rattus norvegicus
Project Data Type: Transcriptome or Gene expression
Attributes: Scope: Multiisolate; Material: Transcriptome; Capture: Whole; Method type: Array
Relevance: Model Organism
Other Accessions: GEO:GSE6943
NAVIGATE ACROSS
1650 additional projects are related by organism.
Project Data:
Resource Name
Number of Links
PUBLICATIONS
PubMed 1
OTHER DATASETS
GEO DataSets 2
GEO Data Details
Parameter
Value
Data volume, Spots 191076
Data volume, Processed Mbytes 4
Data volume, Supplementary Mbytes 17

Select GEO Dataset Browser data

The above **link (http://www.ncbi.nlm.nih.gov/bioproject?Db=gds&DbFrom=bioproject&Cmd=Link&LinkName=bioproject_gds&LinkReadableName=GEO%20DataSets&ordinalpos=1&IdsFromResult=98125)** brings you to the GDS page shown next. The first reference links to GEO2R while the second is annotated with a heatmap and links to GEO Dataset Browser.

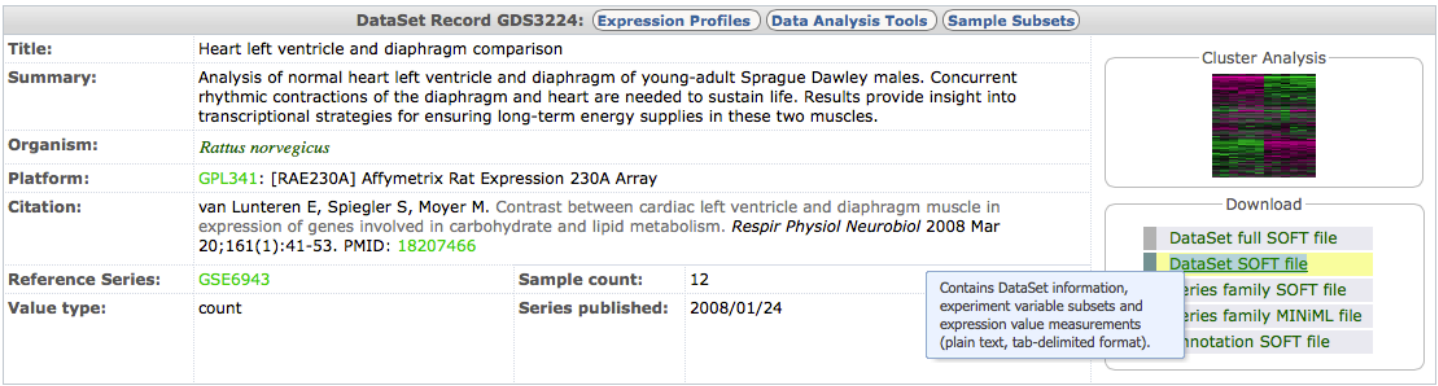

GEO DataSets
GEO DataSets
Advanced
Display Settings: Summary, Sorted by Default order
Send to:
Results: 2
Normal Heart vs Normal Diaphragm
1. (Submitter supplied) Comparison of gene expression of heart (left vent) and diaphragm of normal Sprague Dawley rats, young adult Keywords: Cell type comparison
Organism: Rattus norvegicus
Type: Expression profiling by array
Dataset: GDS3224 Platform: GPL341 12 Samples
Download data: GEO (CEL)
Series Accession: GSE6943 ID: 200006943
PubMed Similar studies Analyze with GEO2R
Heart left ventricle and diaphragm comparison
2. Analysis of normal heart left ventricle and diaphragm of young-adult Sprague Dawley males. Concurrent rhythmic contractions of the diaphragm and heart are needed to sustain life. Results provide insight into transcriptional strategies for ensuring long-term energy supplies in these two muscles.
Organism: Rattus norvegicus
Type: Expression profiling by array, count, 2 tissue sets
Platform: GPL341 Series: GSE6943 12 Samples
Download data: GEO (CEL)
DataSet Accession: GDS3224 ID: 3224
PubMed Similar studies GEO Profiles Analyze DataSet

The highlighted **link (http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3224)** opens in a new tab



Download the full data table

Getting the full table of expression values can be important to extract multiple gene values or any other aim you could have in mind. Tou can get the full dataset from the Download item *on the right of the window*


DataSet Record GDS3224:
Expression Profiles
Data Analysis Tools
Sample Subsets
Title:
Heart left ventricle and diaphragm comparison
Summary:
Analysis of normal heart left ventricle and diaphragm of young-adult Sprague Dawley males. Concurrent rhythmic contractions of the diaphragm and heart are needed to sustain life. Results provide insight into transcriptional strategies for ensuring long-term energy supplies in these two muscles.
Organism:
Rattus norvegicus
Platform:
GPL341: [RAE230A] Affymetrix Rat Expression 230A Array
Citation:
van Lunteren E, Spiegler S, Moyer M. Contrast between cardiac left ventricle and diaphragm muscle in expression of genes involved in carbohydrate and lipid metabolism. Respir Physiol Neurobiol 2008 Mar 20;161(1):41-53. PMID: 18207466
Reference Series:
GSE6943
Sample count:
12
Value type:
count
Series published:
2008/01/24
Cluster Analysis
Download
DataSet full SOFT file
DataSet SOFT file
Contains DataSet information, experiment variable subsets and expression value measurements (plain text, tab-delimited format).

The resulting text file contains a 49 lines header that may pose problems in Excel

split the file in two using some CLI magic *[Collapse]*

```
# download the file from the 'DataSet SOFT file' link in the main window
wget ftp://ftp.ncbi.nlm.nih.gov/geo/datasets/GDS3nnn/GDS3224/soft/GDS3224.soft.gz

# decompress and replace line ends bby valid unix CR
gunzip -c GDS3224.soft.gz | tr "\r" "\n" > GDS3224.soft.txt

# find table line#1
grep -n 'dataset_table_begin' GDS3224.soft.txt
49:!dataset_table_begin

# we need only the number 49!
header_end=$(grep -n 'dataset_table_begin' GDS3224.soft.txt | awk 'BEGIN{FS=":"}{print $1}')

# show the result
echo ${header_end}
49

# split the file in two
head -${header_end} GDS3224.soft.txt > GDS3224.data-header.txt
cat GDS3224.soft.txt | sed -e "1,${header_end}d" > GDS3224.data.txt

# inspect results
grep ^#GSM GDS3224.data-header.txt
#GSM160089 = Value for GSM160089: Diaphragm 1; src: Diaphragm
#GSM160090 = Value for GSM160090: Diaphragm 2; src: Diaphragm
#GSM160091 = Value for GSM160091: Diaphragm 3; src: Diaphragm
#GSM160092 = Value for GSM160092: Diaphragm 4; src: Diaphragm
#GSM160093 = Value for GSM160093: Diaphragm 5; src: Diaphragm
#GSM160094 = Value for GSM160094: Diaphragm 6; src: Diaphragm
#GSM160095 = Value for GSM160095: Heart 1; src: Heart (left vent)
#GSM160096 = Value for GSM160096: Heart 2; src: Heart (left vent)
#GSM160097 = Value for GSM160097: Heart 3; src: Heart (left vent)
#GSM160098 = Value for GSM160098: Heart 4; src: Heart (left vent)
#GSM160099 = Value for GSM160099: Heart 5; src: Heart (left vent)
#GSM160100 = Value for GSM160100: Heart 6; src: Heart (left vent)

head GDS3224.data.txt | column -t
ID_REF      IDENTIFIER  GSM160089  GSM160090  GSM160091  GSM160092  GSM160093  GSM160094  GSM160095  GSM160096  GSM160097  GSM160098  GSM16
1367452_at  Sumo2       2532.9     2518.6     2384.6     2304       2360       2482.8     3166       2938.9     2953.3     2558.8     3043.
1367453_at  Cdc37       3464.2     3197.4     3487.1     3133.2     3432.5     3486.9     3860.2     3429.4     3381.2     4131.3     4364.
1367454_at  Copb2       1620.8     1870.5     1538.6     1334       1502.9     1520.3     1849.8     1852       1858.5     1483.3     1766.
1367455_at  Vcp         5512.5     4103.9     5746.5     4393.6     5870.7     5851.2     5408.8     4682.6     4734.7     4403.7     3940
1367456_at  Ube2d3      6090.8     5352.2     5614.9     5249.6     5834.6     5915.9     3995.3     4356.9     4675.1     4570.4     4994.
```

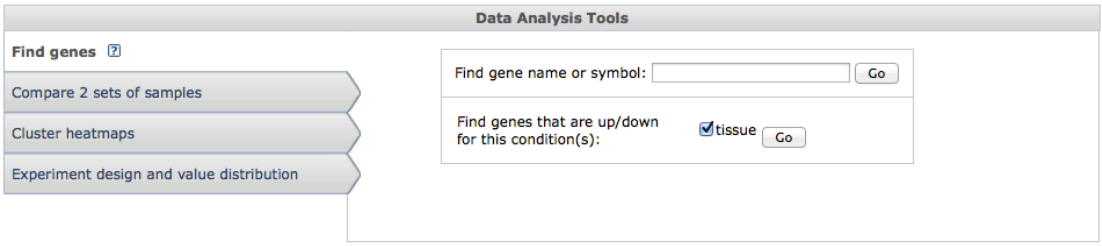
1367457_at	Becn1	1093.9	1134.3	736.4	1219	774.9	712.2	892.1	998.9	782.3	710.8	923.2
1367458_at	Lypla2	347.8	223.9	261.4	338.8	249.6	363.7	422.2	409.9	273.3	492.2	458
1367459_at	Arf1	7665.8	7415.9	7075.9	7349.4	6406.7	6664.6	10400.5	9729.2	9679.2	9996.8	9783.
1367460_at	Gdi2	3155.7	2946.9	3589.7	3487.4	3131.5	3338	3198	2991.3	2781	2754.1	2406.

Walking through the tools

Several built-in tools are ready to serve you on demand. We provide below a rapid overviews of the results obtained by each tool.

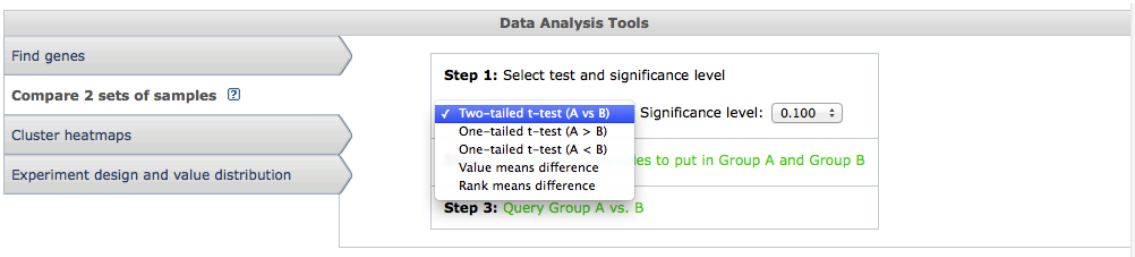
Find Genes

Allows user interested by a particular gene to get its expression values across samples. This simple feature is of limited interest for most applications



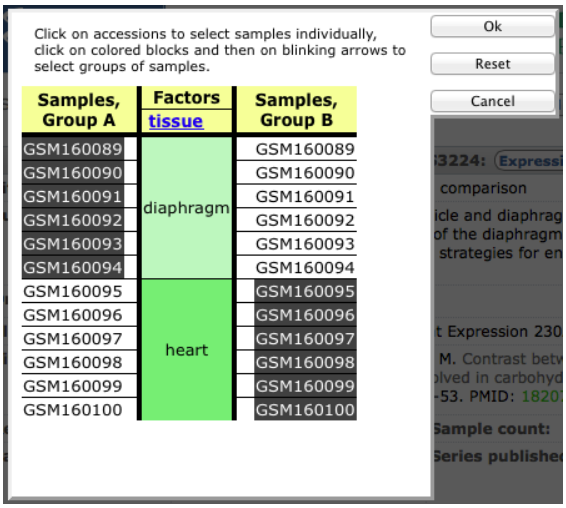
Compare two sets of samples

This allows fine tuning the system and select two groups of samples that are compared using T-test statistics.

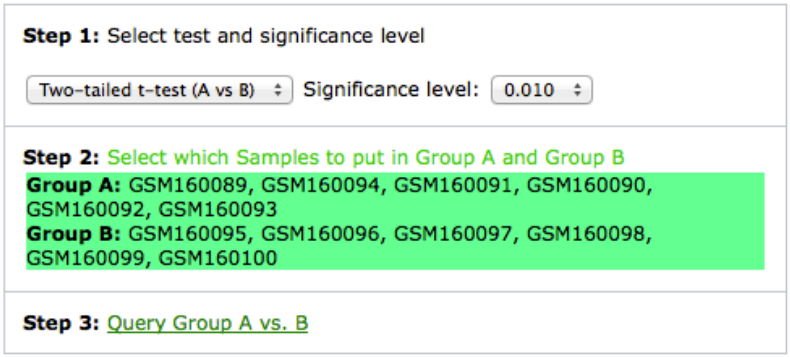


Two tails comparison

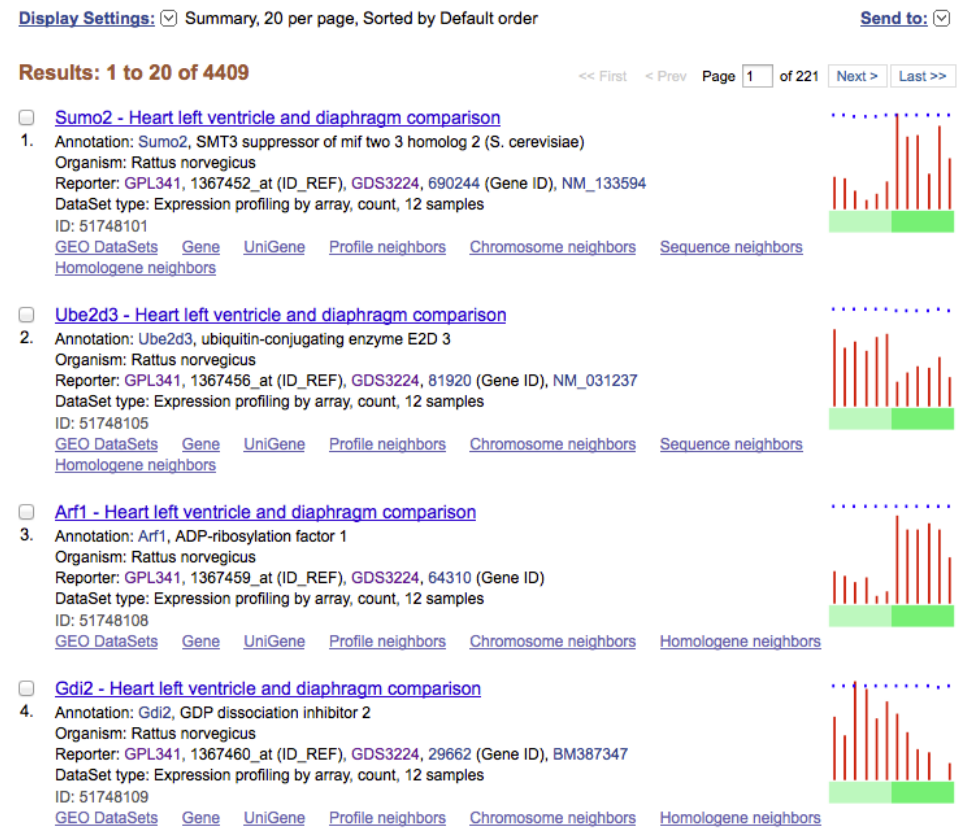
The groups are defined using the mouse.



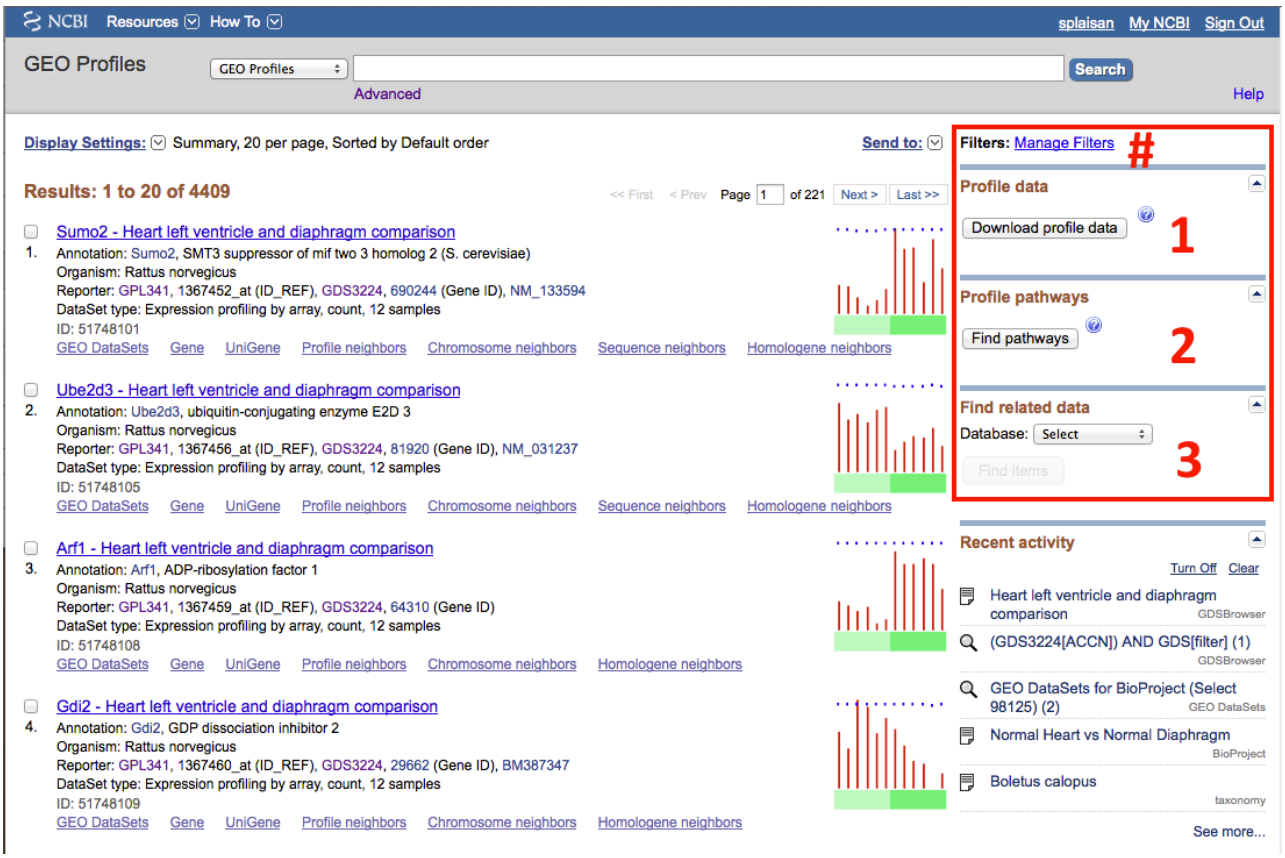
The comparison is based on the selected test method. The choice of the tail(s) allow finding genes less, more or simply differentially expressed.


Step 1: Select test and significance level
Two-tailed t-test (A vs B)
Significance level:
0.010
Step 2: Select which Samples to put in Group A and Group B
Group A: GSM160089, GSM160094, GSM160091, GSM160090, GSM160092, GSM160093
Group B: GSM160095, GSM160096, GSM160097, GSM160098, GSM160099, GSM160100
Step 3: Query Group A vs. B

The result is a very long list of DE genes with barplot view on the right confirming the expression difference but not really useful as-is


Display Settings: Summary, 20 per page, Sorted by Default order
Send to:
Results: 1 to 20 of 4409
<< First < Prev Page 1 of 221 Next > Last >>
1. Sumo2 - Heart left ventricle and diaphragm comparison
Annotation: Sumo2, SMT3 suppressor of mif two 3 homolog 2 (S. cerevisiae)
Organism: Rattus norvegicus
Reporter: GPL341, 1367452_at (ID_REF), GDS3224, 690244 (Gene ID), NM_133594
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748101
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors
Homologene neighbors
2. Ube2d3 - Heart left ventricle and diaphragm comparison
Annotation: Ube2d3, ubiquitin-conjugating enzyme E2D 3
Organism: Rattus norvegicus
Reporter: GPL341, 1367456_at (ID_REF), GDS3224, 81920 (Gene ID), NM_031237
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748105
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors
Homologene neighbors
3. Arf1 - Heart left ventricle and diaphragm comparison
Annotation: Arf1, ADP-ribosylation factor 1
Organism: Rattus norvegicus
Reporter: GPL341, 1367459_at (ID_REF), GDS3224, 64310 (Gene ID)
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748108
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors
4. Gdi2 - Heart left ventricle and diaphragm comparison
Annotation: Gdi2, GDP dissociation inhibitor 2
Organism: Rattus norvegicus
Reporter: GPL341, 1367460_at (ID_REF), GDS3224, 29662 (Gene ID), BM387347
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748109
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors

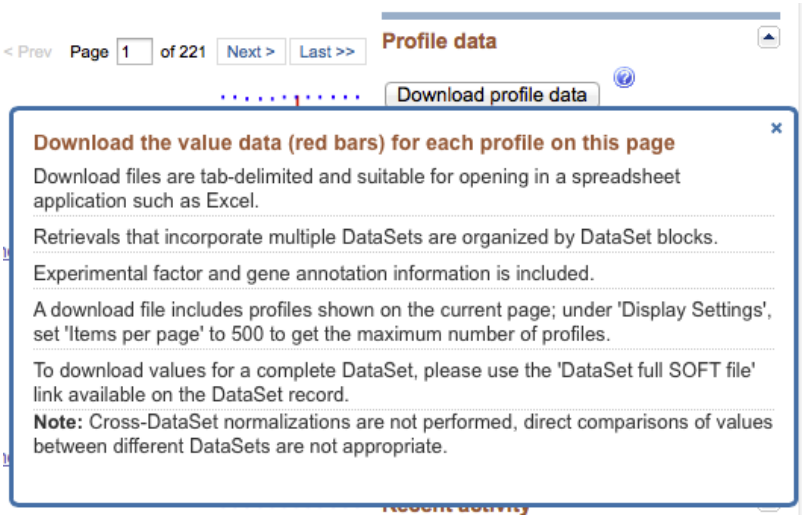
Luckily, this is not all of it and right menus allow doing more with the found items

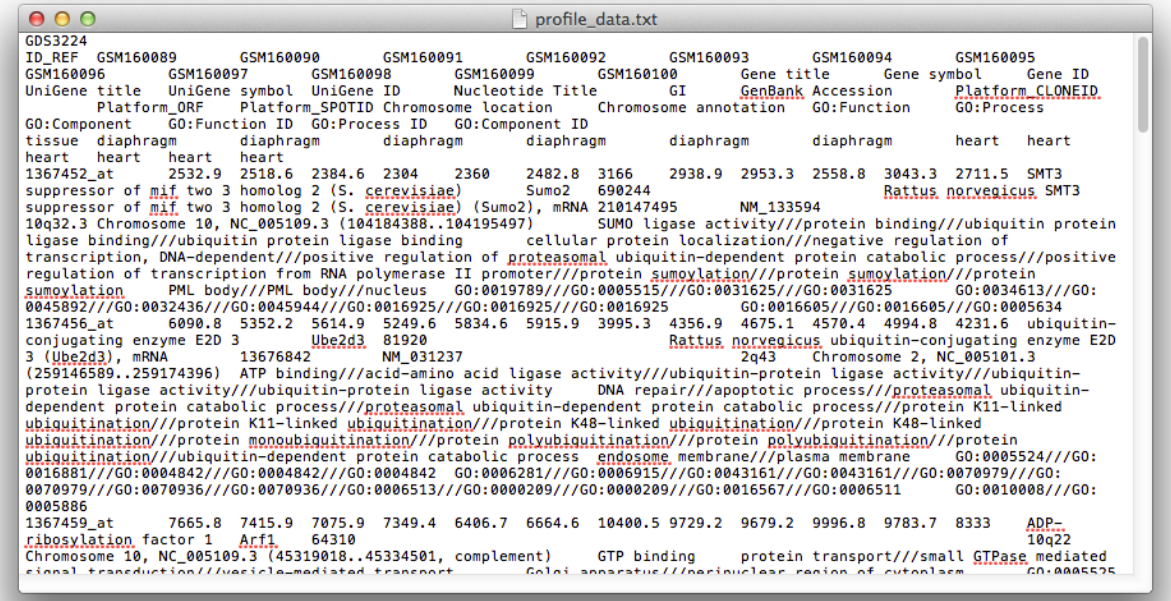

NCBI Resources How To
splaisan My NCBI Sign Out
GEO Profiles
GEO Profiles
Search
Advanced
Help
Display Settings: Summary, 20 per page, Sorted by Default order
Send to:
Results: 1 to 20 of 4409
<< First < Prev Page 1 of 221 Next > Last >>
1. Sumo2 - Heart left ventricle and diaphragm comparison
Annotation: Sumo2, SMT3 suppressor of mif two 3 homolog 2 (S. cerevisiae)
Organism: Rattus norvegicus
Reporter: GPL341, 1367452_at (ID_REF), GDS3224, 690244 (Gene ID), NM_133594
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748101
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors
2. Ube2d3 - Heart left ventricle and diaphragm comparison
Annotation: Ube2d3, ubiquitin-conjugating enzyme E2D 3
Organism: Rattus norvegicus
Reporter: GPL341, 1367456_at (ID_REF), GDS3224, 81920 (Gene ID), NM_031237
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748105
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors
3. Arf1 - Heart left ventricle and diaphragm comparison
Annotation: Arf1, ADP-ribosylation factor 1
Organism: Rattus norvegicus
Reporter: GPL341, 1367459_at (ID_REF), GDS3224, 64310 (Gene ID)
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748108
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors
4. Gdi2 - Heart left ventricle and diaphragm comparison
Annotation: Gdi2, GDP dissociation inhibitor 2
Organism: Rattus norvegicus
Reporter: GPL341, 1367460_at (ID_REF), GDS3224, 29662 (Gene ID), BM387347
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748109
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors
Filters: Manage Filters
#
Profile data
Download profile data
1
Profile pathways
Find pathways
2
Find related data
Database: Select
Find items
3
Recent activity
Turn Off Clear
Heart left ventricle and diaphragm comparison GDSBrowser
(GDS3224[ACCN]) AND GDS[filter] (1) GDSBrowser
GEO DataSets for BioProject (Select 98125) (2) GEO DataSets
Normal Heart vs Normal Diaphragm BioProject
Boletus calopus taxonomy
See more...

Filters *and* ***Find related data*** *are not detailed here and left to your curiosity*

Profile data

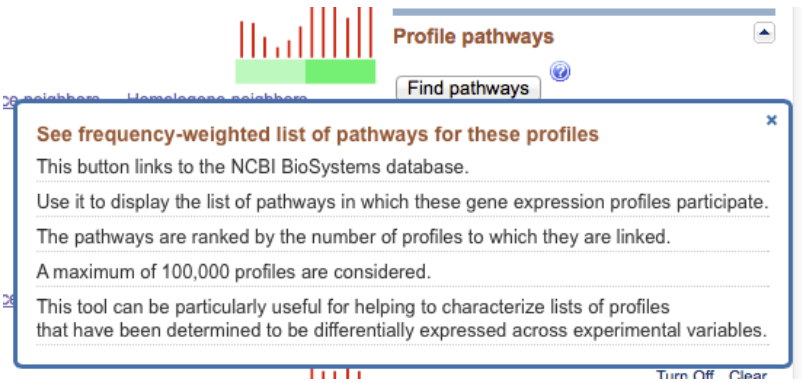
This lets you download differentially expressed data selected by the tool

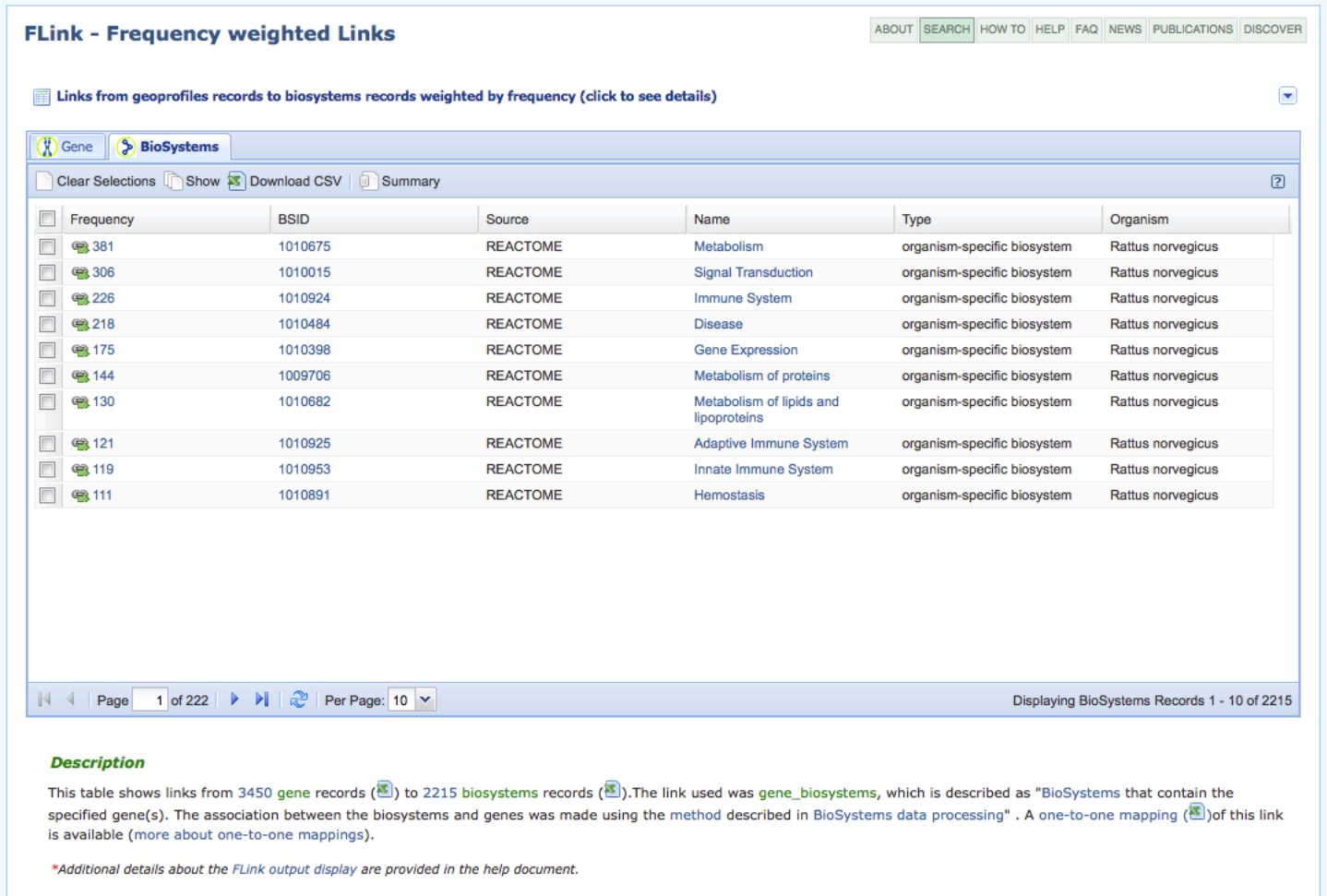

< Prev
Page 1 of 221
Next >
Last >>
Profile data
Download profile data
Download the value data (red bars) for each profile on this page
Download files are tab-delimited and suitable for opening in a spreadsheet application such as Excel.
Retrievals that incorporate multiple DataSets are organized by DataSet blocks.
Experimental factor and gene annotation information is included.
A download file includes profiles shown on the current page; under 'Display Settings', set 'Items per page' to 500 to get the maximum number of profiles.
To download values for a complete DataSet, please use the 'DataSet full SOFT file' link available on the DataSet record.
Note: Cross-DataSet normalizations are not performed, direct comparisons of values between different DataSets are not appropriate.


profile_data.txt
GDS3224
ID_REF GSM160089 GSM160090 GSM160091 GSM160092 GSM160093 GSM160094 GSM160095
GSM160096 GSM160097 GSM160098 GSM160099 GSM160100 Gene title Gene symbol Gene ID
UniGene title UniGene symbol UniGene ID Nucleotide Title GI GenBank Accession Platform_CLONEID
Platform_ORF Platform_SPOTID Chromosome location Chromosome annotation GO:Function GO:Process
GO:Component GO:Function ID GO:Process ID GO:Component ID
tissue diaphragm diaphragm diaphragm diaphragm diaphragm diaphragm heart heart
heart heart heart heart
1367452_at 2532.9 2518.6 2384.6 2304 2360 2482.8 3166 2938.9 2953.3 2558.8 3043.3 2711.5 SMT3
suppressor of mif two 3 homolog 2 (S. cerevisiae) Sumo2 690244 Rattus norvegicus SMT3
suppressor of mif two 3 homolog 2 (S. cerevisiae) (Sumo2), mRNA 210147495 NM_133594
10q32.3 Chromosome 10, NC_005109.3 (104184388..104195497) SUMO ligase activity///protein binding///ubiquitin protein
ligase binding///ubiquitin protein ligase binding cellular protein localization///negative regulation of
transcription, DNA-dependent///positive regulation of proteasomal ubiquitin-dependent protein catabolic process///positive
regulation of transcription from RNA polymerase II promoter///protein sumoylation///protein sumoylation///protein
sumoylation PML body///PML body///nucleus GO:0019789///GO:0005515///GO:0031625///GO:0031625 GO:0034613///GO:
0045892///GO:0032436///GO:0045944///GO:0016925///GO:0016925///GO:0016925 GO:0016605///GO:0016605///GO:0005634
1367456_at 6090.8 5352.2 5614.9 5249.6 5834.6 5915.9 3995.3 4356.9 4675.1 4570.4 4994.8 4231.6 ubiquitin-
conjugating enzyme E2D 3 Ube2d3 81920 Rattus norvegicus ubiquitin-conjugating enzyme E2D
3 (Ube2d3), mRNA 13676842 NM_031237 2q43 Chromosome 2, NC_005101.3
(259146589..259174396) ATP binding///acid-amino acid ligase activity///ubiquitin-protein ligase activity///ubiquitin-
protein ligase activity///ubiquitin-protein ligase activity DNA repair///apoptotic process///proteasomal ubiquitin-
dependent protein catabolic process///proteasomal ubiquitin-dependent protein catabolic process///protein K11-linked
ubiquitination///protein K11-linked ubiquitination///protein K48-linked ubiquitination///protein K48-linked
ubiquitination///protein monoubiquitination///protein polyubiquitination///protein polyubiquitination///protein
ubiquitination///ubiquitin-dependent protein catabolic process endosome membrane///plasma membrane GO:0005524///GO:
0016881///GO:0004842///GO:0004842///GO:0004842 GO:0006281///GO:0006915///GO:0043161///GO:0043161///GO:0070979///GO:
0070979///GO:0070936///GO:0070936///GO:0006513///GO:0000209///GO:0000209///GO:0016567///GO:0006511 GO:0010008///GO:
0005886
1367459_at 7665.8 7415.9 7075.9 7349.4 6406.7 6664.6 10400.5 9729.2 9679.2 9996.8 9783.7 8333 ADP-
ribosylation factor 1 Arf1 64310 10q22
Chromosome 10, NC_005109.3 (45319018..45334501, complement) GTP binding protein transport///small GTPase mediated

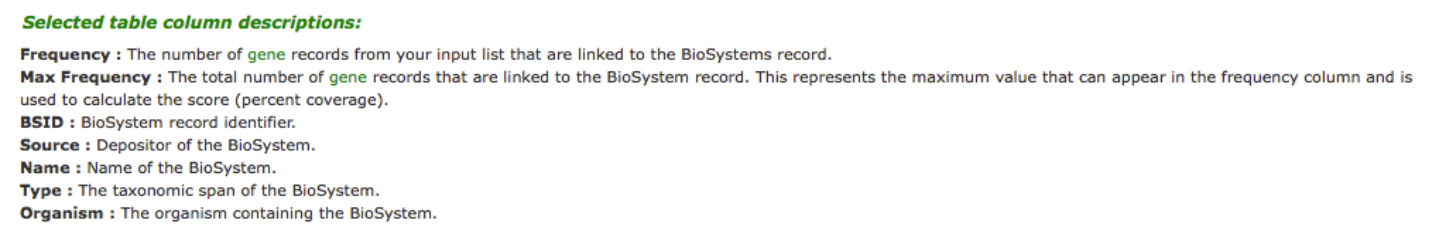
Profile pathways

Search pathways enriched in the obtained subset


Profile pathways
Find pathways
See frequency-weighted list of pathways for these profiles
This button links to the NCBI BioSystems database.
Use it to display the list of pathways in which these gene expression profiles participate.
The pathways are ranked by the number of profiles to which they are linked.
A maximum of 100,000 profiles are considered.
This tool can be particularly useful for helping to characterize lists of profiles that have been determined to be differentially expressed across experimental variables.


FLink - Frequency weighted Links
ABOUT SEARCH HOW TO HELP FAQ NEWS PUBLICATIONS DISCOVER
Links from geoprofiles records to biosystems records weighted by frequency (click to see details)
Gene BioSystems
Clear Selections Show Download CSV Summary
Frequency BSID Source Name Type Organism
381 1010675 REACTOME Metabolism organism-specific biosystem Rattus norvegicus
306 1010015 REACTOME Signal Transduction organism-specific biosystem Rattus norvegicus
226 1010924 REACTOME Immune System organism-specific biosystem Rattus norvegicus
218 1010484 REACTOME Disease organism-specific biosystem Rattus norvegicus
175 1010398 REACTOME Gene Expression organism-specific biosystem Rattus norvegicus
144 1009706 REACTOME Metabolism of proteins organism-specific biosystem Rattus norvegicus
130 1010682 REACTOME Metabolism of lipids and lipoproteins organism-specific biosystem Rattus norvegicus
121 1010925 REACTOME Adaptive Immune System organism-specific biosystem Rattus norvegicus
119 1010953 REACTOME Innate Immune System organism-specific biosystem Rattus norvegicus
111 1010891 REACTOME Hemostasis organism-specific biosystem Rattus norvegicus
Page 1 of 222 Per Page: 10
Displaying BioSystems Records 1 - 10 of 2215
Description
This table shows links from 3450 gene records ( ) to 2215 biosystems records ( ).The link used was gene_biosystems, which is described as "BioSystems that contain the specified gene(s). The association between the biosystems and genes was made using the method described in BioSystems data processing" . A one-to-one mapping ( )of this link is available (more about one-to-one mappings).
*Additional details about the FLink output display are provided in the help document.

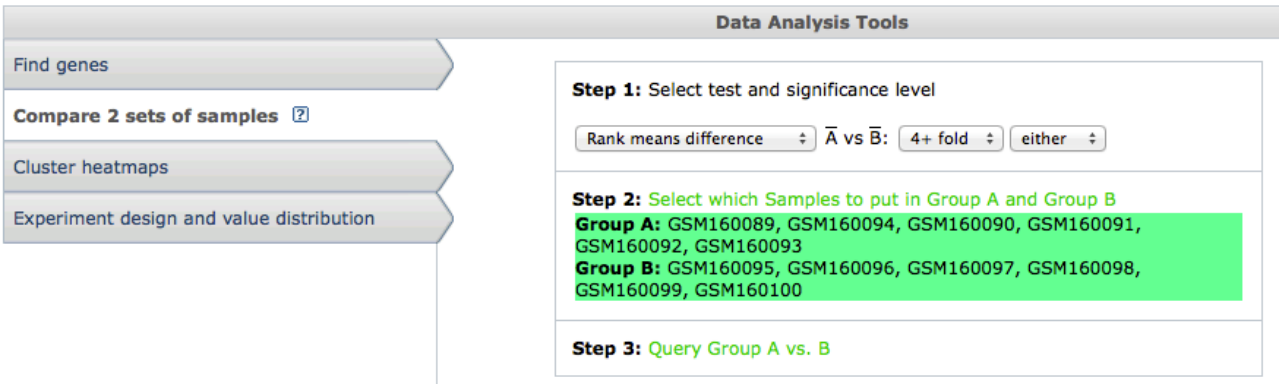
Column legends

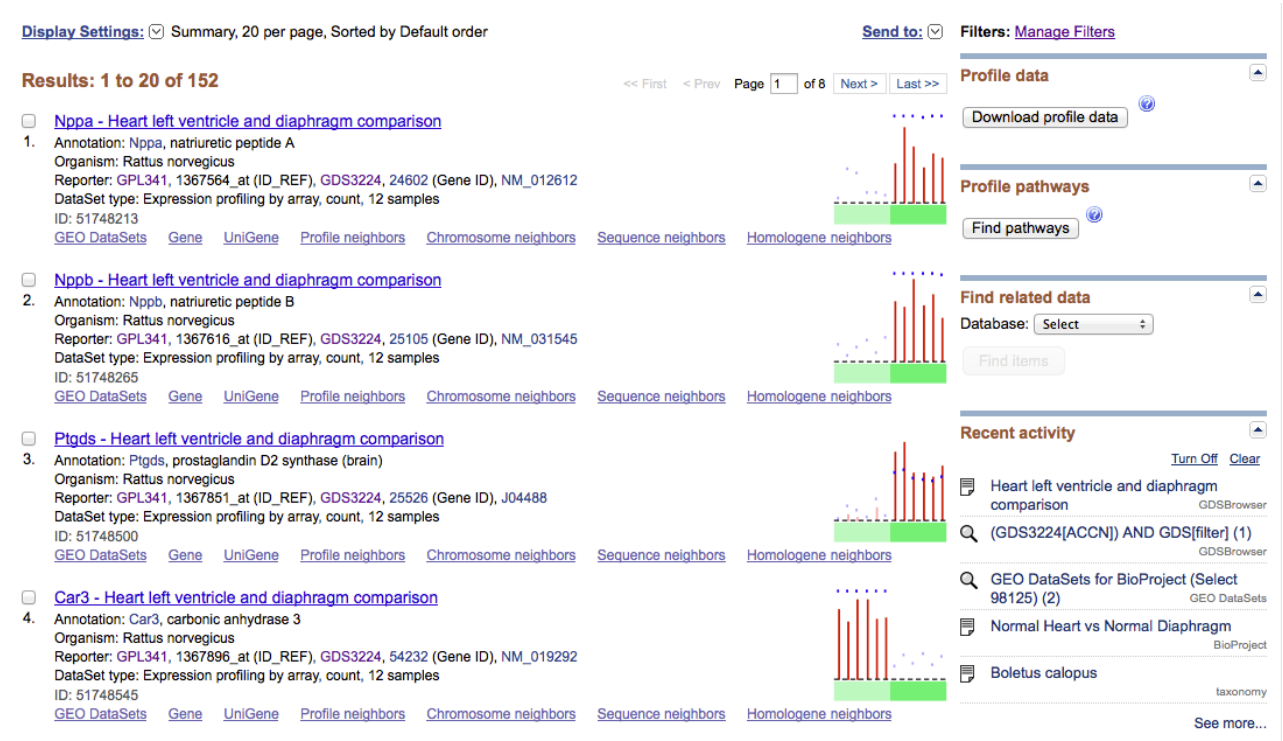

Selected table column descriptions:
Frequency : The number of gene records from your input list that are linked to the BioSystems record.
Max Frequency : The total number of gene records that are linked to the BioSystem record. This represents the maximum value that can appear in the frequency column and is used to calculate the score (percent coverage).
BSID : BioSystem record identifier.
Source : Depositor of the BioSystem.
Name : Name of the BioSystem.
Type : The taxonomic span of the BioSystem.
Organism : The organism containing the BioSystem.

Rank mean difference "4x & either"

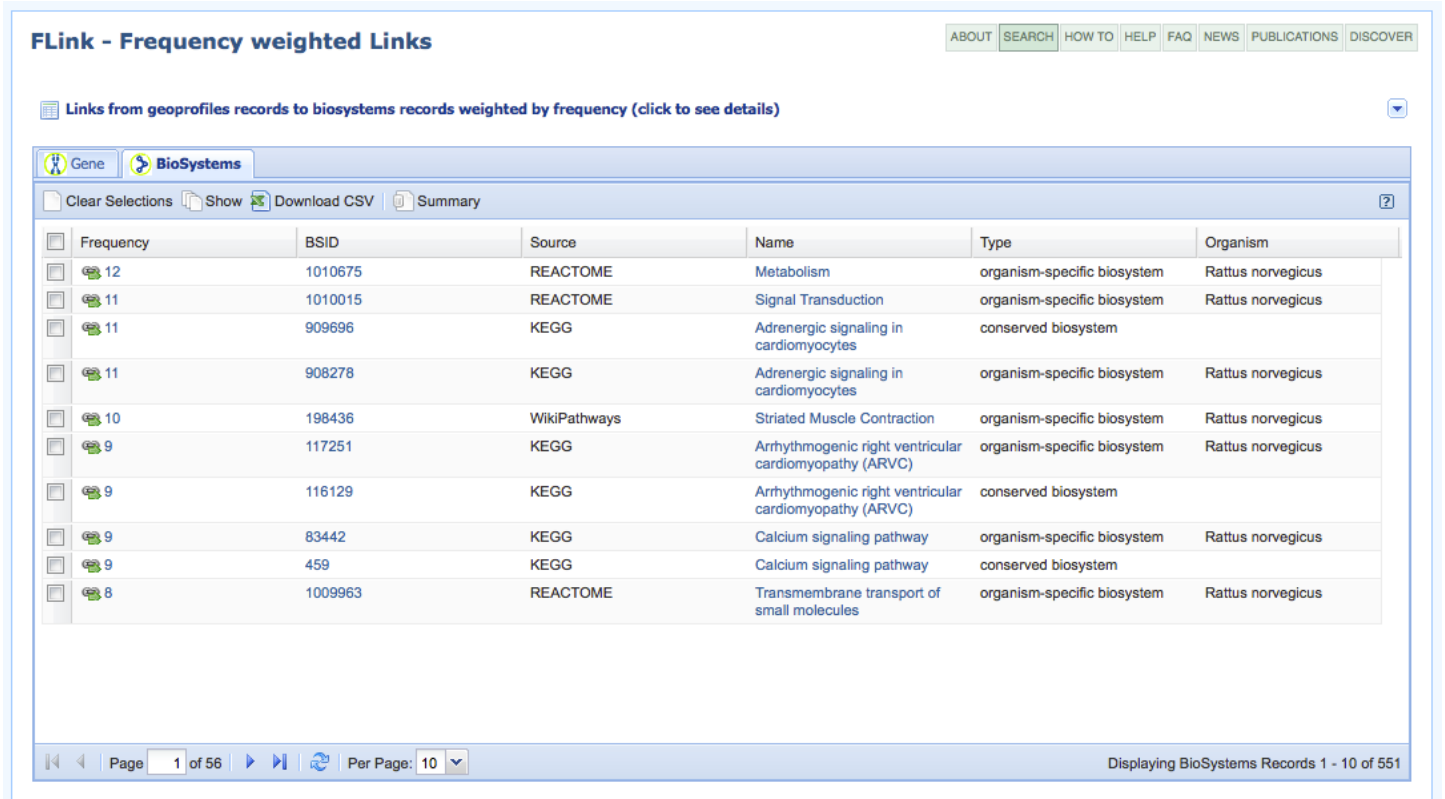
The former test found too many hits (n=4409) to be specific for given pathways, if we try instead the **rank mean difference** test with a 4-fold difference cutoff, we get only **152** genes that are probably more specific for the biology behind the sample groups.

Search pathways enriched in the obtained subset


Data Analysis Tools
Find genes
Compare 2 sets of samples
Cluster heatmaps
Experiment design and value distribution
Step 1: Select test and significance level
Rank means difference
A vs B:
4+ fold
either
Step 2: Select which Samples to put in Group A and Group B
Group A: GSM160089, GSM160094, GSM160090, GSM160091, GSM160092, GSM160093
Group B: GSM160095, GSM160096, GSM160097, GSM160098, GSM160099, GSM160100
Step 3: Query Group A vs. B


Display Settings: Summary, 20 per page, Sorted by Default order
Send to:
Filters: Manage Filters
Results: 1 to 20 of 152
<< First < Prev Page 1 of 8 Next > Last >>
1. Nppa - Heart left ventricle and diaphragm comparison
Annotation: Nppa, natriuretic peptide A
Organism: Rattus norvegicus
Reporter: GPL341, 1367564_at (ID_REF), GDS3224, 24602 (Gene ID), NM_012612
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748213
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors
2. Nppb - Heart left ventricle and diaphragm comparison
Annotation: Nppb, natriuretic peptide B
Organism: Rattus norvegicus
Reporter: GPL341, 1367616_at (ID_REF), GDS3224, 25105 (Gene ID), NM_031545
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748265
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors
3. Ptgds - Heart left ventricle and diaphragm comparison
Annotation: Ptgds, prostaglandin D2 synthase (brain)
Organism: Rattus norvegicus
Reporter: GPL341, 1367851_at (ID_REF), GDS3224, 25526 (Gene ID), J04488
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748500
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors
4. Car3 - Heart left ventricle and diaphragm comparison
Annotation: Car3, carbonic anhydrase 3
Organism: Rattus norvegicus
Reporter: GPL341, 1367896_at (ID_REF), GDS3224, 54232 (Gene ID), NM_019292
DataSet type: Expression profiling by array, count, 12 samples
ID: 51748545
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors
Profile data
Download profile data
Profile pathways
Find pathways
Find related data
Database: Select
Find items
Recent activity
Turn Off Clear
Heart left ventricle and diaphragm comparison GDSBrowser
(GDS3224[ACCN]) AND GDS[filter] (1) GDSBrowser
GEO DataSets for BioProject (Select 98125) (2) GEO DataSets
Normal Heart vs Normal Diaphragm BioProject
Boletus calopus taxonomy
See more...

The pathways are now very specific to heart biology as expected

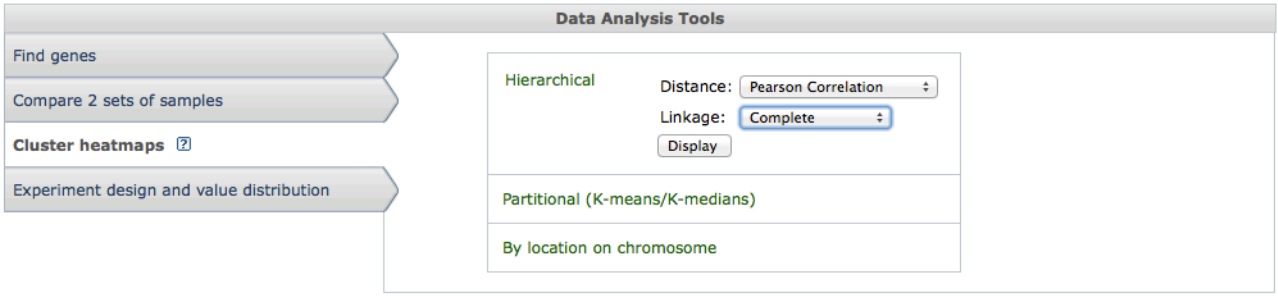

FLink - Frequency weighted Links
ABOUT SEARCH HOW TO HELP FAQ NEWS PUBLICATIONS DISCOVER
Links from geoprofiles records to biosystems records weighted by frequency (click to see details)
Gene BioSystems
Clear Selections Show Download CSV Summary
Frequency BSID Source Name Type Organism
12 1010675 REACTOME Metabolism organism-specific biosystem Rattus norvegicus
11 1010015 REACTOME Signal Transduction organism-specific biosystem Rattus norvegicus
11 909696 KEGG Adrenergic signaling in cardiomyocytes conserved biosystem
11 908278 KEGG Adrenergic signaling in cardiomyocytes organism-specific biosystem Rattus norvegicus
10 198436 WikiPathways Striated Muscle Contraction organism-specific biosystem Rattus norvegicus
9 117251 KEGG Arrhythmogenic right ventricular cardiomyopathy (ARVC) organism-specific biosystem Rattus norvegicus
9 116129 KEGG Arrhythmogenic right ventricular cardiomyopathy (ARVC) conserved biosystem
9 83442 KEGG Calcium signaling pathway organism-specific biosystem Rattus norvegicus
9 459 KEGG Calcium signaling pathway conserved biosystem
8 1009963 REACTOME Transmembrane transport of small molecules organism-specific biosystem Rattus norvegicus
Page 1 of 56 Per Page: 10
Displaying BioSystems Records 1 - 10 of 551

Cluster heatmaps

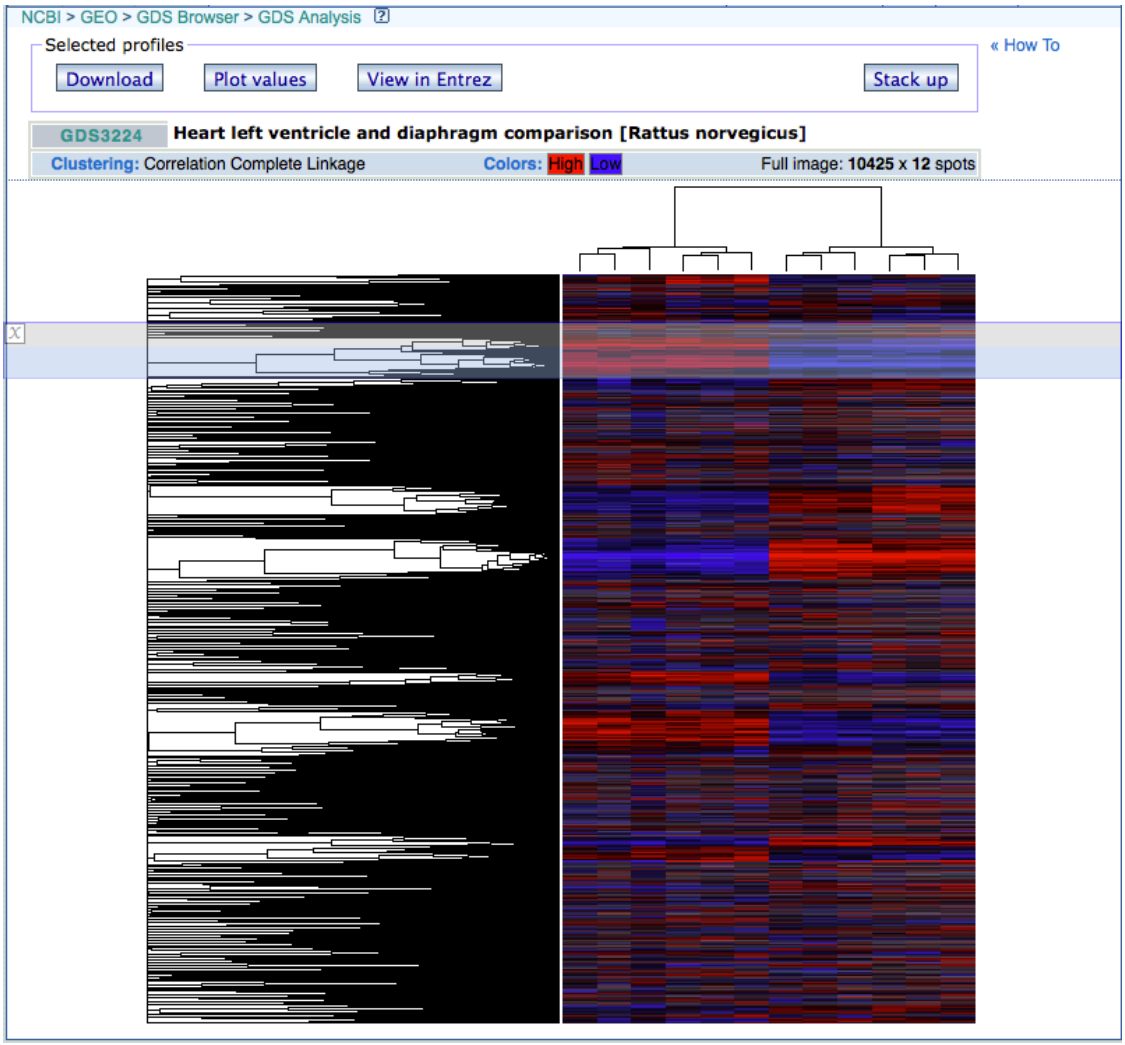
Two options are presented in this part, the hierarchical clustering and the KMean approaches.

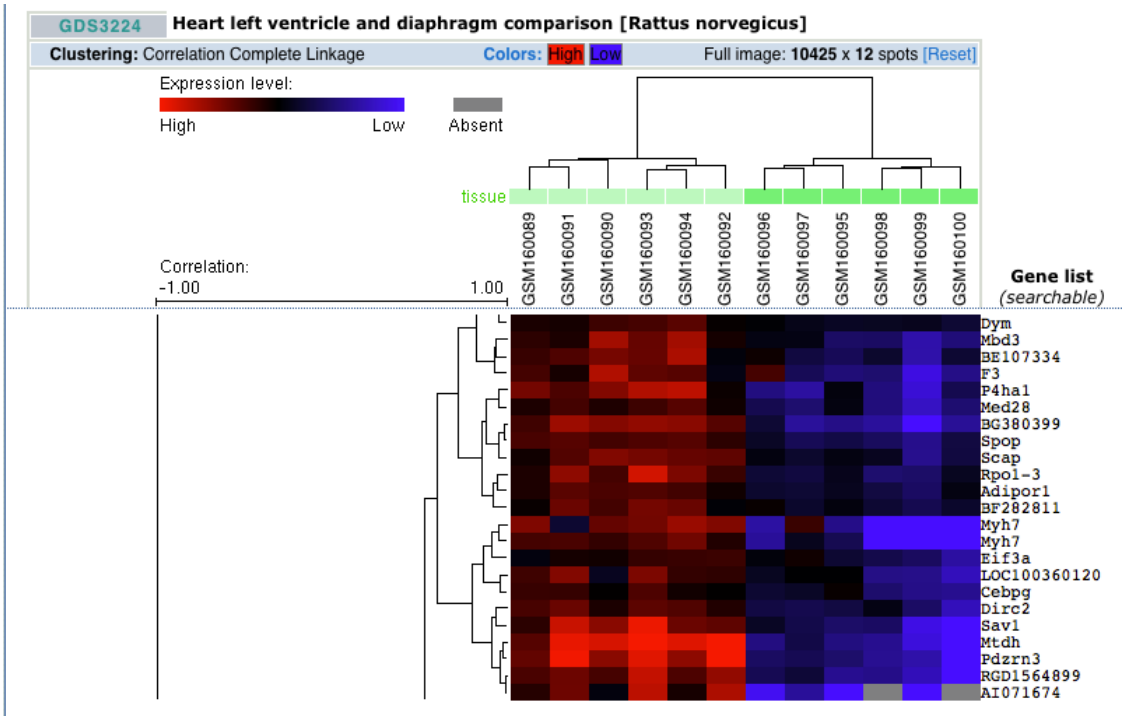
Hierarchical clustering

Hierarchical clustering and heatmaps are classical representations of differential expression that nicely provide high level view on the data. Both are done for you by the browser and return images as well as the corresponding tables for further exploration and without one line of code.



A number of interactive links allow changing the correlation method, changing colors, and/or selecting heatmap regions to plot other graphs or export data to file



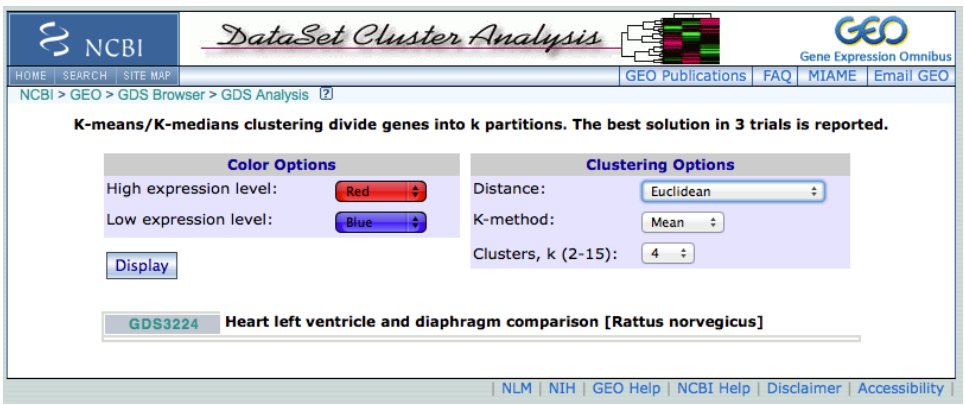


KMean clustering

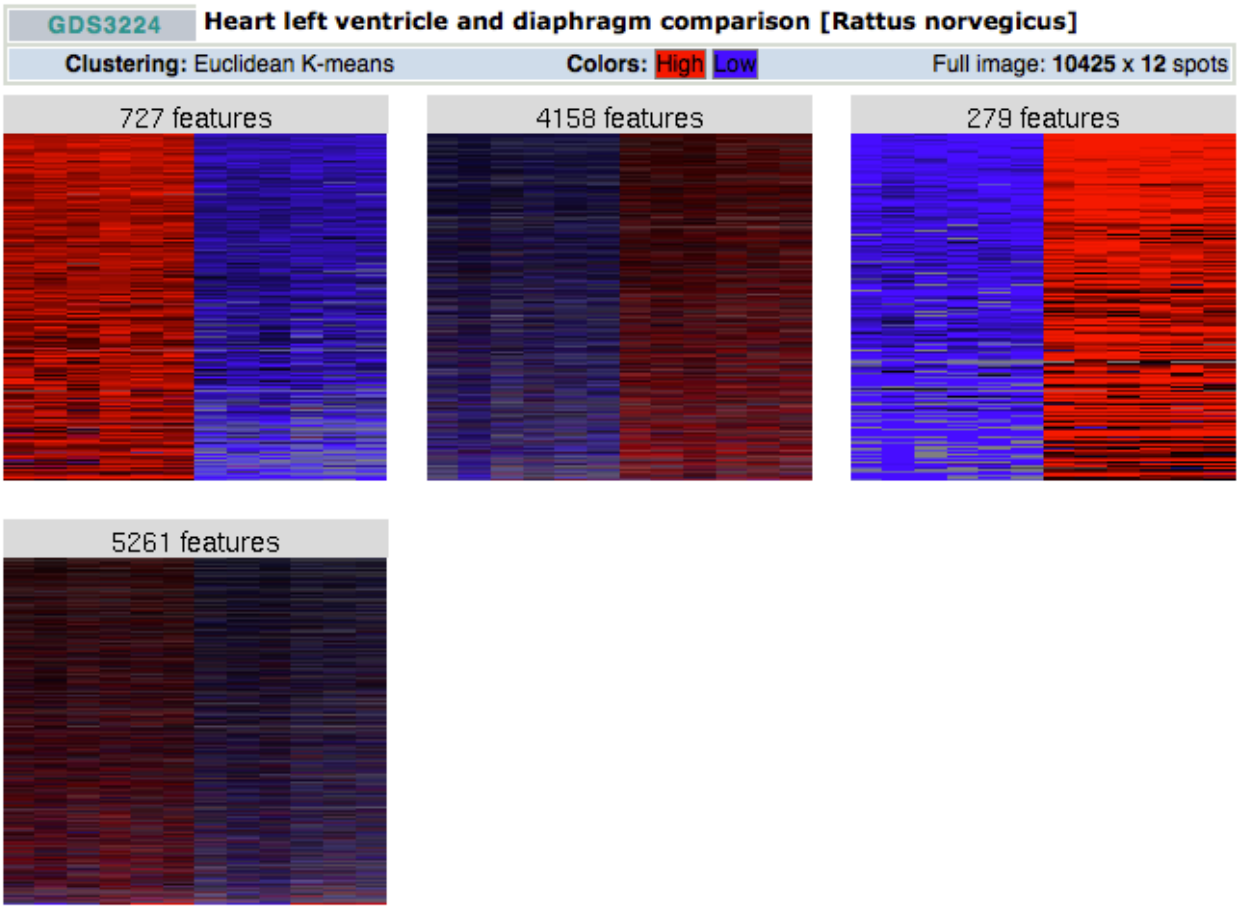
KMean 'biclusters' are un-supervized results showing groups of genes and samples that behave in a very similar way. This method offers the advantage of not being biased by knowledge and only directed by the data itself.

The trade-off is that running Kmean over and over again will always return results and not necessarily the same results. This seems counterintuitive for the average a Biologist who is trained towards reproducibility but is a reality for the analyst ***you will not get the same results as the ones below!!! BUT you will find back mostly the same genes in similar clusters***

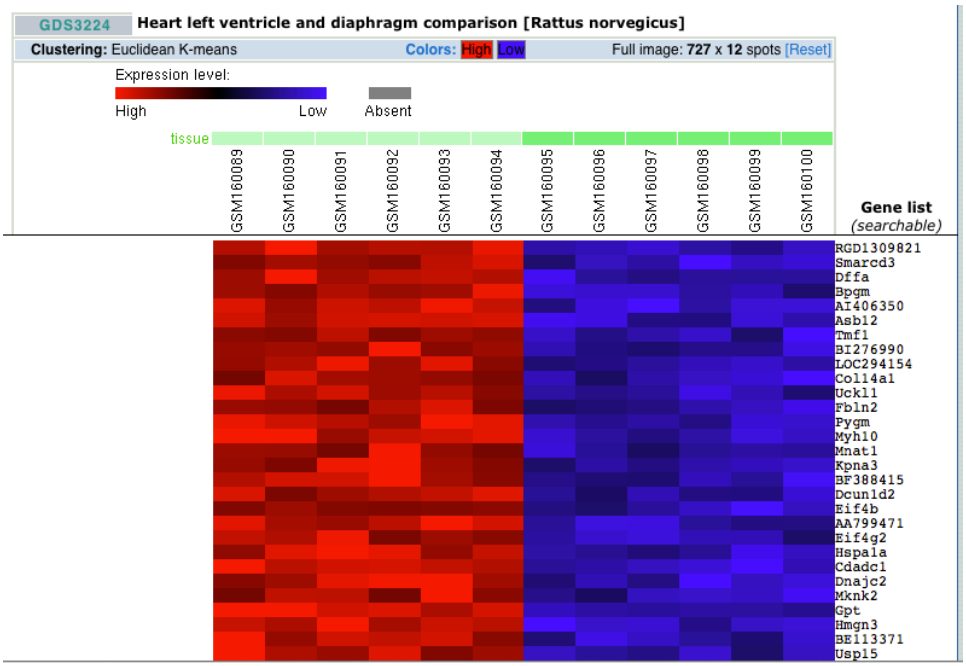
KMean clusters can identify core processes represented in sample groups by a small set of genes whose co-regulation is very clear. The user decides arbitrarily of the number of clusters to be found. Too few will lead to mixed profiles while too many will lead to apparent redundancy.



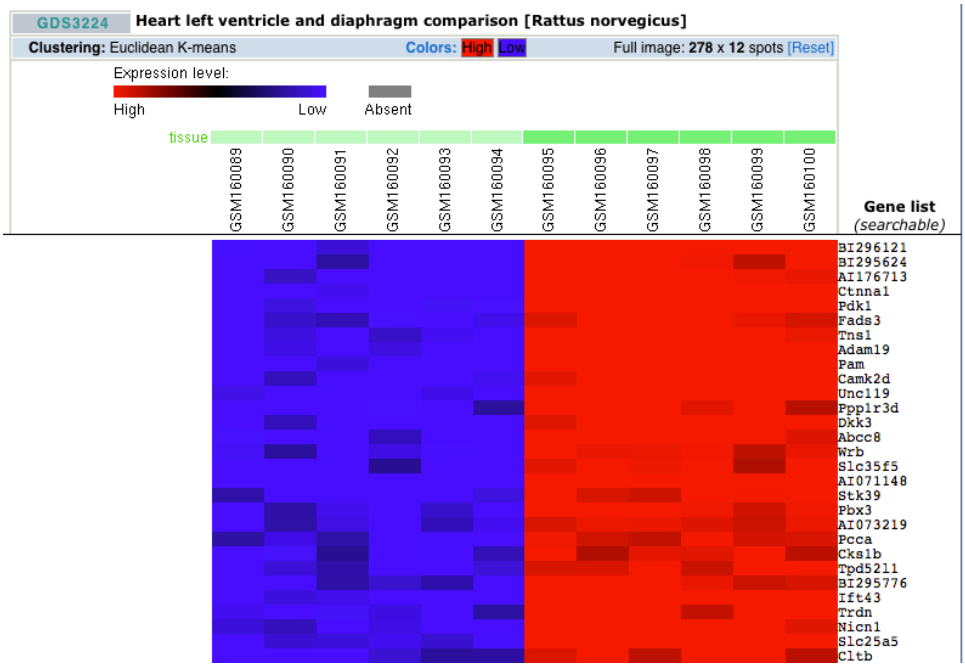
From the obtained solution, it is clear that the four clusters are grouping genes having clear trends in both sample groups. Groups #1 and #3 being highly contrasted between Diaphragm and Heart while #2 and #4 are more subtle DE groups



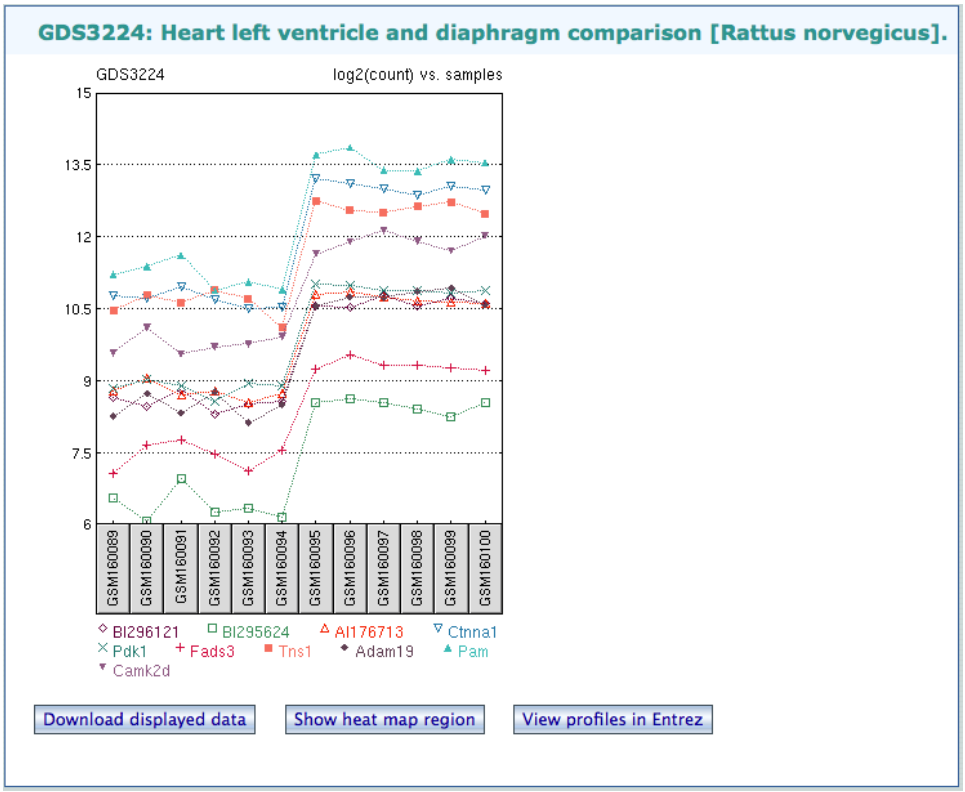
Clicking on the first group to get more details for genes higher in the Diaphragm



Similarly, for genes higher in the Heart



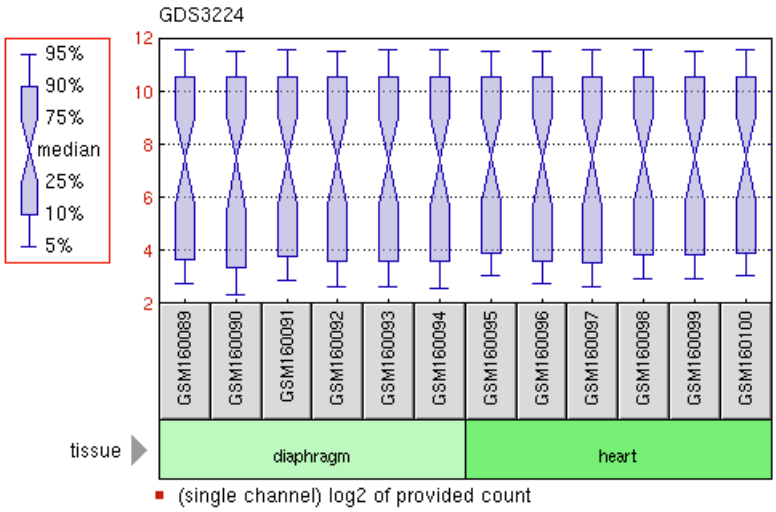
Finally, zooming on a small number of genes allows plotting their profiles and exporting the results to file



Experiment design and value distribution

This QC tool plots box-plot for each sample and allows evaluating the global quality of the dataset. A good dataset has a constant median line and similar distributions.

Profile GDS3224
Title Heart left ventricle and diaphragm comparison
Organism Rattus norvegicus



Graph caption help

Sample	Title
GSM160089	Diaphragm 1
GSM160090	Diaphragm 2
GSM160091	Diaphragm 3
GSM160092	Diaphragm 4
GSM160093	Diaphragm 5
GSM160094	Diaphragm 6
GSM160095	Heart 1
GSM160096	Heart 2
GSM160097	Heart 3
GSM160098	Heart 4
GSM160099	Heart 5
GSM160100	Heart 6

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑ http://www.ncbi.nlm.nih.gov/sites/GDSbrowser/
2. ↑ http://www.ncbi.nlm.nih.gov/geo/info/datasets.html

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=PubMA_Exercise.3&oldid=10946"
Category: PUBMA2014

- This page was last modified on 26 August 2014, at 13:39.
- This page has been accessed 160 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

PubMA Exercise.4

From BioWareWIKI

RobiNA analysis

Contents

- 1 Introduction
- 2 Required before starting RobiNA
  - 2.1 RobiNA working great ... or NOT
  - 2.2 The CDF annotation file
- 3 Step by Step RobiNA analysis workflow
  - 3.1 start RobiNA and create a new project for results
  - 3.2 Importing CDF & CEL files
  - 3.3 Performing QC on each CEL file
    - 3.3.1 Evaluating QC results
  - 3.4 Define design
  - 3.5 Computing differential expression
    - 3.5.1 Reviewing DE results
- 4 Adding annotations to the RobiNA data
- 5 Conclusion
- 6 download exercise files

Introduction

" Several commercial programs such as GeneSpring featuring a rich and simple user interface to statistically analyze high-throughput "omics" data are available. However, these are usually very expensive and might even require an annual subscription. On the other hand, there are free, open source tools such as BioConductor which offer great statistical options and support for free, but this power often comes at the price of usability, since these tools often feature either only a command line interface or solely very basic user interaction.

Finally, there are tools such as RMA Express which offer a rich user interface but only a very limited set of options.

RobiNA tries to bridge this gap by providing a flexible user friendly graphical interface to unleash the power of R/BioConductor for the individual biologist. RobiNA comes as a convenient all-in-one installation package, that automatically installs the application itself plus all required external tools (i.e. the R and BioConductor frameworks and bowtie). "[1]

Although RobiNA can handle several data type including two color microarray and even RNASeq data, we here provide a simple tutorial to perform QC and differential analysis of Affymetrix microarray data comparable to what can be achieved using the CLC Main workbench

Please see the **RobiNA quick guide (http://mapman.gabipd.org/c/document_library/get_file?uuid=a09272e9-e474-402e-a554-b03d6ec9efd6&groupId=10207)** and **Robin and RobiNA user's manual (http://mapman.gabipd.org/c/document_library/get_file?uuid=60912d03-660e-4281-9834-22f2789424d2&groupId=10207)** which contain step-by-step walk through and detailed information. RobinA can be downloaded from **http://mapman.gabipd.org/web/guest/robin**

Required before starting RobiNA

RobiNA working great ... or NOT

*RobiNA should work under Windows, Mac, and Unix as it uses Java which is a universal language. However, we have found that some computers and operating systems are refractory to RobiNA and even with 6GB RAM may have issues running RobiNA with as few as 10 CEL files. RobiNA requires a recent version of Java JDK (you can obtain JDK from: **[1] (http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html)**). The RobiNA developers do not actively support the software right now and you will need to try things by yourself if you have issues*

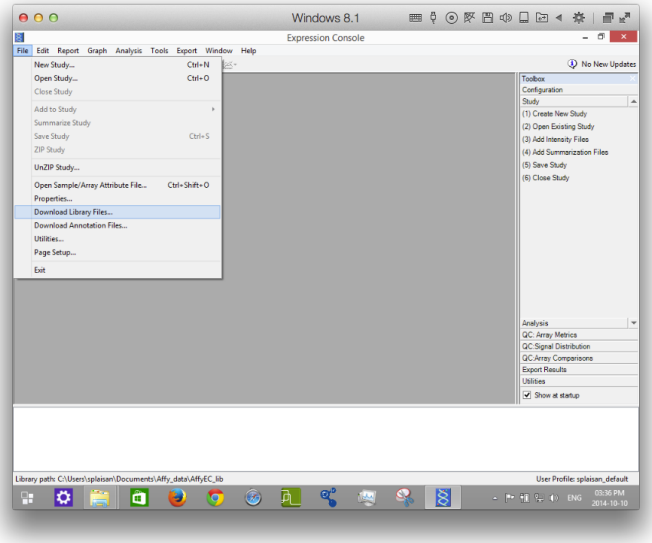
The CDF annotation file

RobiNA needs a **CDF** file to work. CDF files are complex text tables allowing the identification of the probes and genes associated with the chip spots, such files are sometimes difficult to find. The RobinA software was developed by a plant groups and has plant as main focus and does not include non-plant annotations. For those who want to analyze other living organisms, the microarray annotation file corresponding to the used chip (CDF) should be obtained before starting RobiNA. A place where to find Affymetrix CDF files should be the company website but it is often difficult to locate the CDF among the long lists of available Affymetrix resources (**http://www.affymetrix.com/estore/ | free registration required**)[2].

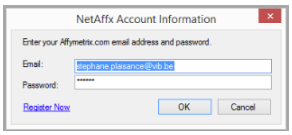
The easiest way to obtain the correct Affymetrix CDF file is probably by using the **Affymetrix Expression Console**.

*the Affymetrix Expression Console software is **only available for Windows** and downloading/using it requires a free user registration*

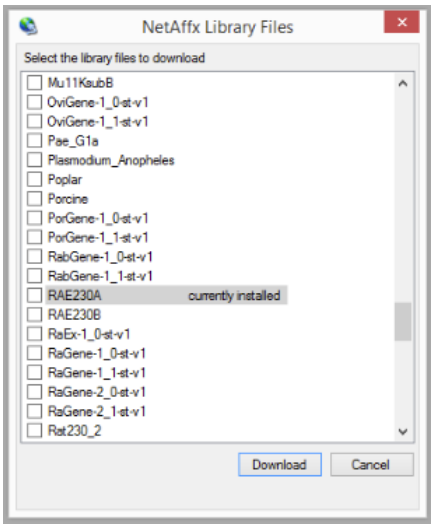
After installing and starting the **Affymetrix Expression Console** program, Use the **File** menu to access the 'library' download manager



Log into the Affy support area using your credentials

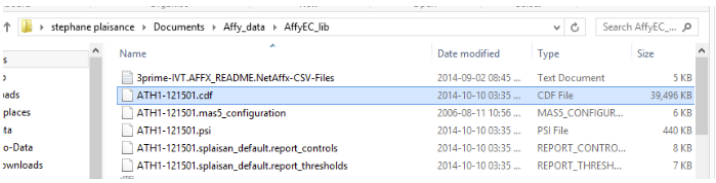


Search and check the chip corresponding to your data, and let the manager download the associated files

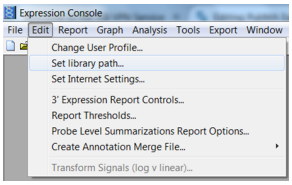


In the example above we highlight the rat array that was already downloaded and was used to generate the demo data in the CLC Bio Main Workbench (measurements of gene expression in tissues from cardiac left ventricle and diaphragm muscle of rats; Lunteren et al., 2008).

The Affymetrix Expression Console will download the CDF file to the folder that you have specified in the library path (typically, the folder where your CEL files are stored). The file shown here as example is for Arabidopsis


stephane plaisance ▸ Documents ▸ Affy_data ▸ AffyEC_lib
Search AffyEC_...
Name
Date modified
Type
Size
3prime-IVT.AFFX_README.NetAffx-CSV-Files 2014-09-02 08:45 ... Text Document 5 KB
ATH1-121501.cdf 2014-10-10 03:35 ... CDF File 39,496 KB
ATH1-121501.mas5_configuration 2006-08-11 10:56 ... MAS5_CONFIGUR... 6 KB
ATH1-121501.psi 2014-10-10 03:35 ... PSI File 440 KB
ATH1-121501.splaisan_default.report_controls 2014-10-10 03:35 ... REPORT_CONTRO... 8 KB
ATH1-121501.splaisan_default.report_thresholds 2014-10-10 03:35 ... REPORT_THRESH... 7 KB

You can change this folder via **Edit** in the top menu. Click **Set library path**.



Step by Step RobiNA analysis workflow

You can use RobiNA for...

- quality assessment of your data
- normalization of your microarray data

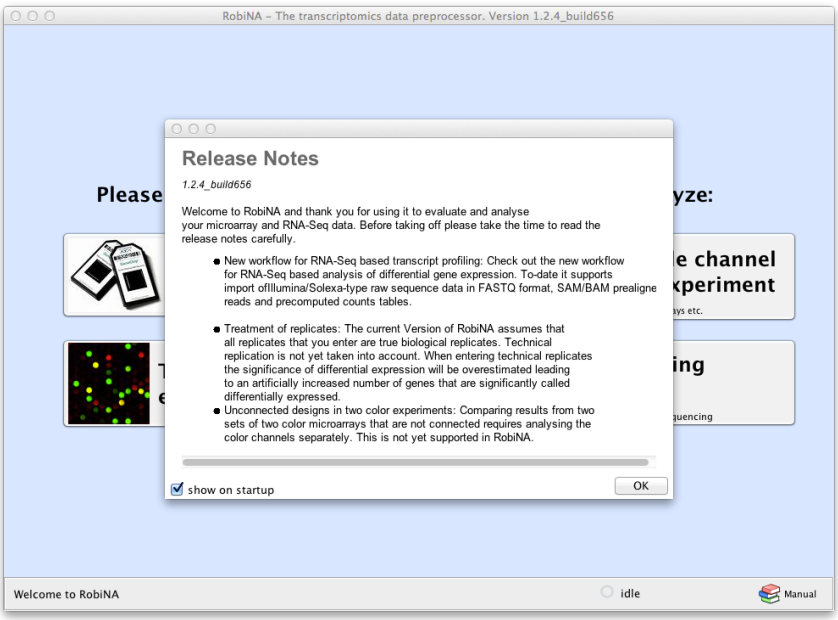
- detection of differentially expressed genes
- preparation of the data for an import into MapMan and/or excel
- generation of informative plots on your experiment

Pre-requisites are the CEL files containing each hybridization data (obtained from the GEO page as a '.tar' archive and decompressed to individual CEL files) and a CDF file listing all probes present on the chip used in the experiment. Today's CDF file can be obtained from the Affymetrix site [2] after registering (free) and searching for the library file corresponding to the platform reported in the GEO pages (in our case: **Rat Expression Set 230, aka RAE230A**).

- http://www.affymetrix.com/Auth/support/downloads/library_files/rae230_libraryfile.zip.

The decompressed archive contains the required CDF file under 'CD_RAE230_rev04/Full/RAE230A/LibFiles/RAE230A.CDF' that can be copied in the RobiNA project area.

start RobiNA and create a new project for results

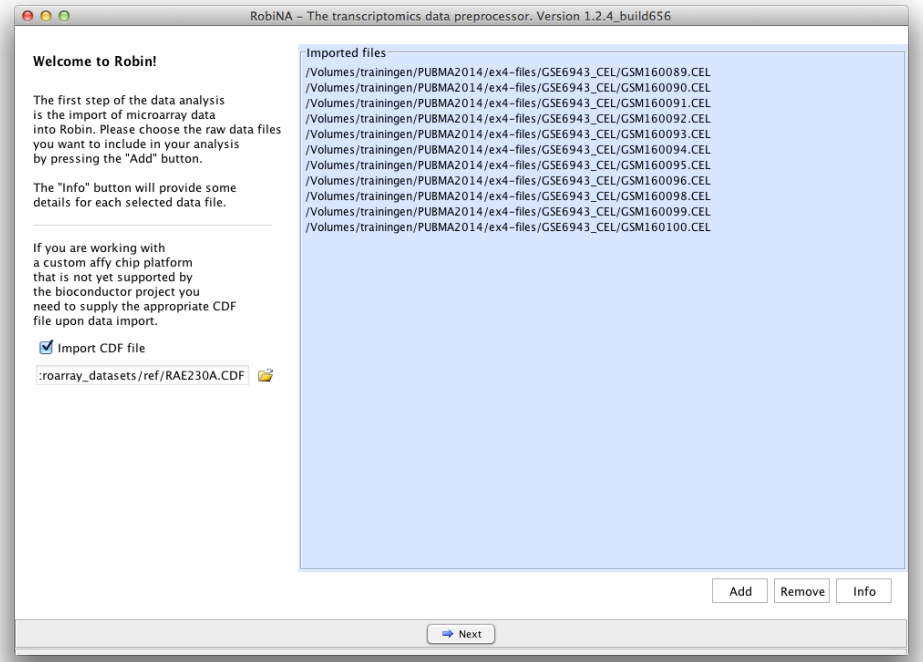

RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Release Notes
1.2.4_build656
Welcome to RobiNA and thank you for using it to evaluate and analyse your microarray and RNA-Seq data. Before taking off please take the time to read the release notes carefully.
New workflow for RNA-Seq based transcript profiling: Check out the new workflow for RNA-Seq based analysis of differential gene expression. To-date it supports import ofIllumina/Solexa-type raw sequence data in FASTQ format, SAM/BAM prealigne reads and precomputed counts tables.
Treatment of replicates: The current Version of RobiNA assumes that all replicates that you enter are true biological replicates. Technical replication is not yet taken into account. When entering technical replicates the significance of differential expression will be overestimated leading to an artificially increased number of genes that are significantly called differentially expressed.
Unconnected designs in two color experiments: Comparing results from two sets of two color microarrays that are not connected requires analysing the color channels separately. This is not yet supported in RobiNA.
show on startup
OK
Welcome to RobiNA
idle
Manual

Welcome to RobiNA
Start new project
Open existing project
The first step of the workflow will beto choose a project directory in which all files related to
the analysis will be stored. Please make sure that the chosen target directory will be on a volume (hard drive, USB stick etc.) that provides enough free space to possibly accomodate your raw data since Robin will make working copies of the data in case processing of the raw files is necessary prior to analysis
Project folder
Free space on target volume:
Cancel
Continue

Importing CDF & CEL files

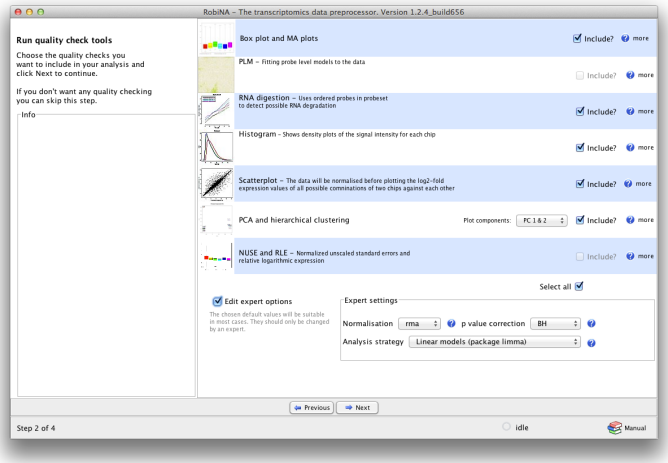
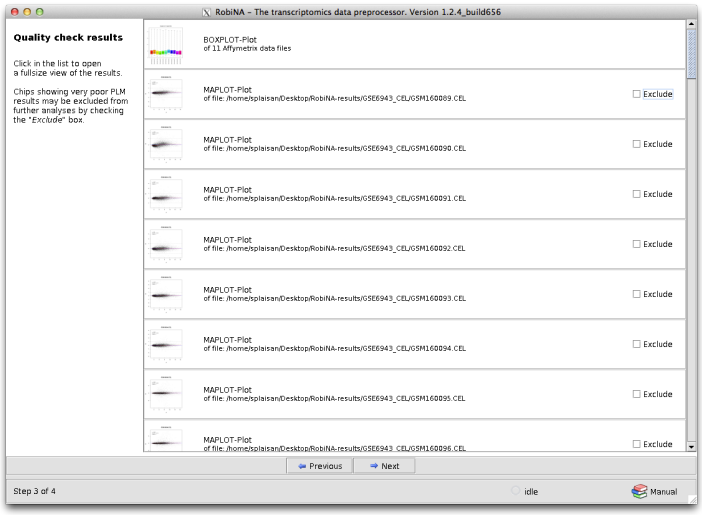
While preparing this training, we discovered that GEO had a damaged file (GSM160097) for one of the sample, we will therefore do this training with only 5-replicate in the Diaphragm group while the CLC analysis was done with the full data

You are now ready to import the CEL files and the matching CDF annotation database.


RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Welcome to Robin!
The first step of the data analysis is the import of microarray data into Robin. Please choose the raw data files you want to include in your analysis by pressing the "Add" button.
The "Info" button will provide some details for each selected data file.
If you are working with a custom affy chip platform that is not yet supported by the bioconductor project you need to supply the appropriate CDF file upon data import.
Import CDF file
:roarray_datasets/ref/RAE230A.CDF
Imported files
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160089.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160090.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160091.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160092.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160093.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160094.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160095.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160096.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160098.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160099.CEL
/Volumes/trainingen/PUBMA2014/ex4-files/GSE6943_CEL/GSM160100.CEL
Add
Remove
Info
Next

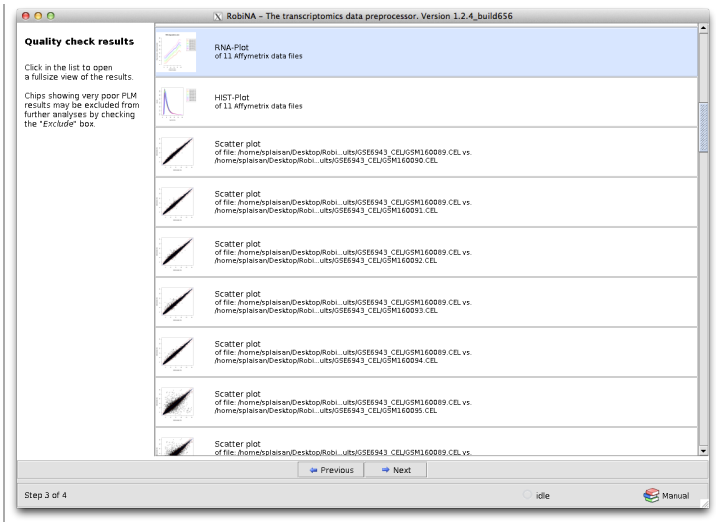
Performing QC on each CEL file

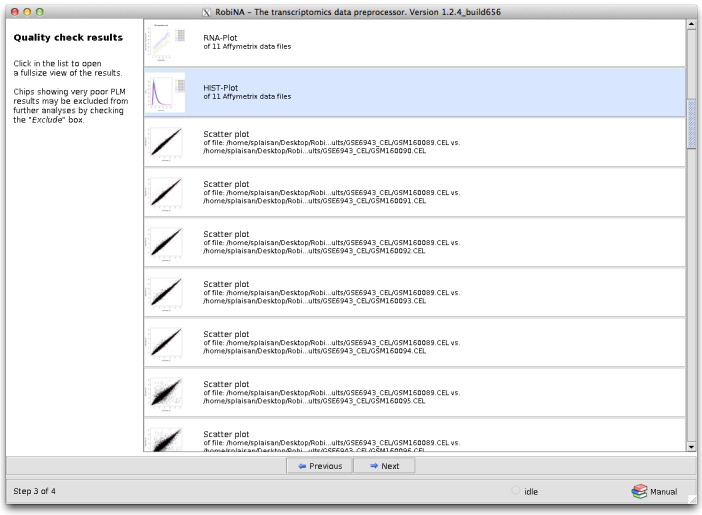
When QC finishes, a number of single plots can be reviewed by clicking on each section. These figures report issues or demonstrate good quality of the data and are also saved to disk for later inclusion in your reports.

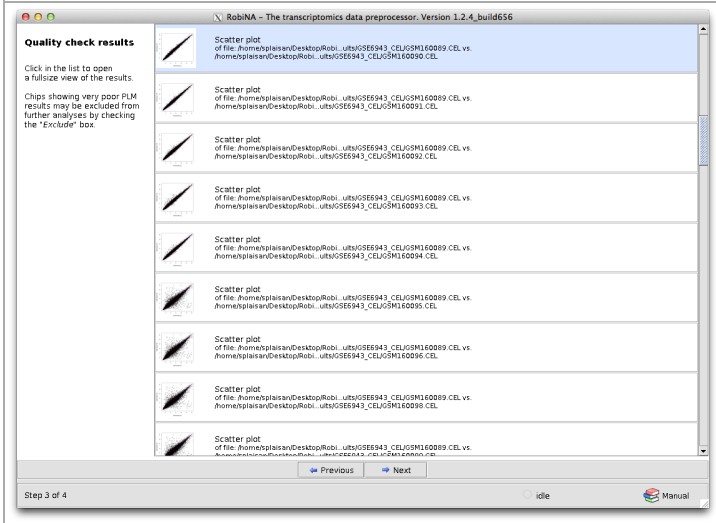
Overview of all QC-Plots	[Collapse]
	

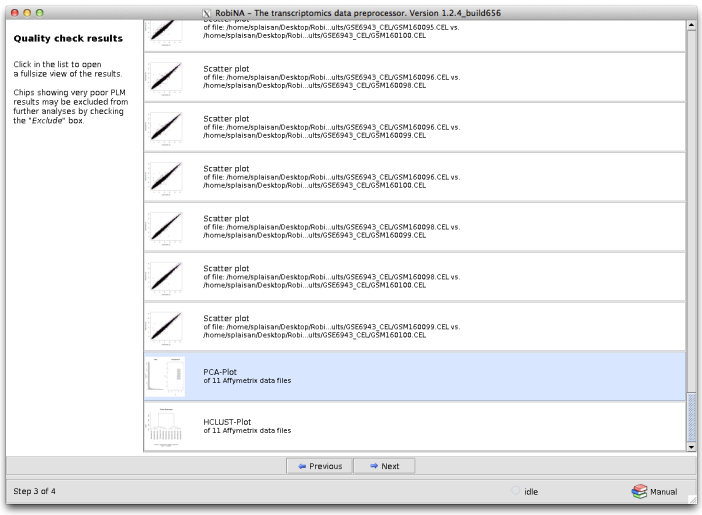
RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Run quality check tools
Choose the quality checks you want to include in your analysis and click Next to continue.
If you don't want any quality checking you can skip this step.
Info
Box plot and MA plots
PLM – Fitting probe level models to the data
RNA digestion – Uses ordered probes in probeset to detect possible RNA degradation
Histogram – Shows density plots of the signal intensity for each chip
Scatterplot – The data will be normalised before plotting the log2-fold expression values of all possible combinations of two chips against each other
PCA and hierarchical clustering
NUSE and RLE – Normalized unscaled standard errors and relative logarithmic expression
Include?
more
Select all
Edit expert options
Expert settings
Normalisation
rma
p value correction
BH
Analysis strategy
Linear models (package limma)
Previous
Next
Step 2 of 4
idle
Manual

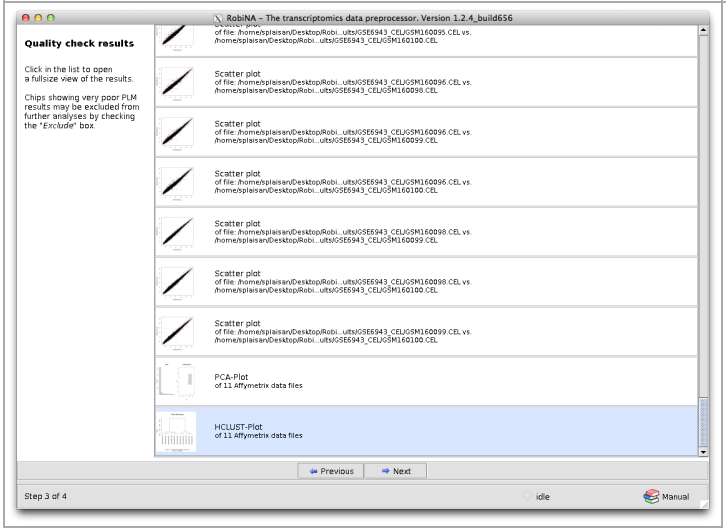
RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Quality check results
Click in the list to open a fullsize view of the results.
Chips showing very poor PLM results may be excluded from further analyses by checking the "Exclude" box.
BOXPLOT-Plot
of 11 Affymetrix data files
MAPLOT-Plot
Exclude
Previous
Next
Step 3 of 4
idle
Manual


RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Quality check results
Click in the list to open a fullsize view of the results.
Chips showing very poor PLM results may be excluded from further analyses by checking the "Exclude" box.
RNA-Plot
of 11 Affymetrix data files
HIST-Plot
of 11 Affymetrix data files
Scatter plot
Previous
Next
Step 3 of 4
idle
Manual


RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Quality check results
Click in the list to open a fullsize view of the results.
Chips showing very poor PLM results may be excluded from further analyses by checking the "Exclude" box.
RNA-Plot
of 11 Affymetrix data files
HIST-Plot
of 11 Affymetrix data files
Scatter plot
Previous
Next
Step 3 of 4
idle
Manual


RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Quality check results
Click in the list to open a fullsize view of the results.
Chips showing very poor PLM results may be excluded from further analyses by checking the "Exclude" box.
Scatter plot
Previous
Next
Step 3 of 4
idle
Manual

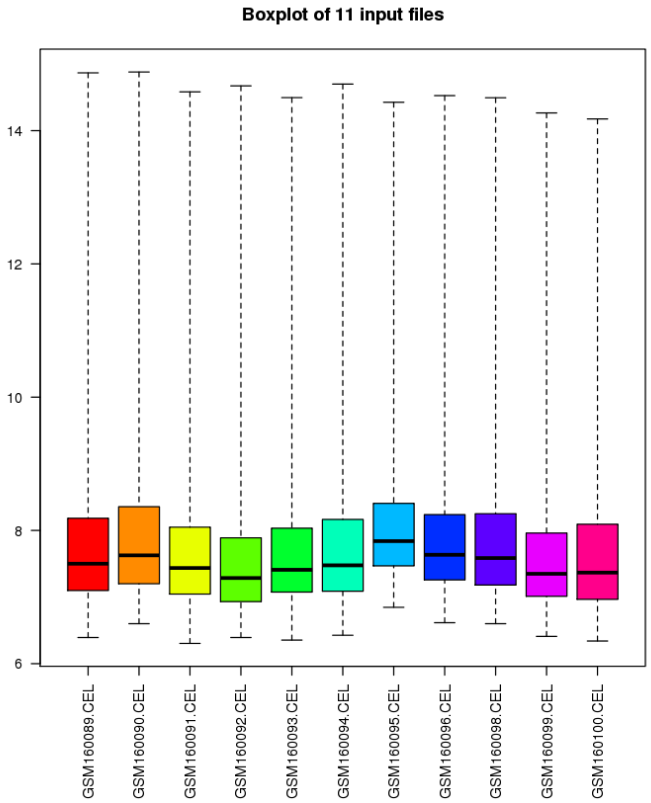

RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Quality check results
Click in the list to open a fullsize view of the results.
Chips showing very poor PLM results may be excluded from further analyses by checking the "Exclude" box.
Scatter plot
PCA-Plot
of 11 Affymetrix data files
HCLUST-Plot
of 11 Affymetrix data files
Previous
Next
Step 3 of 4
idle
Manual

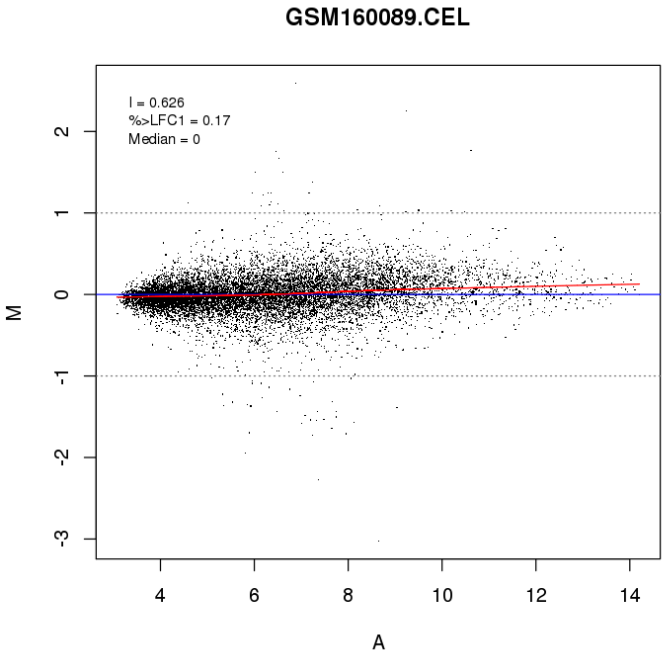

RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Quality check results
Click in the list to open a fullsize view of the results.
Chips showing very poor PLM results may be excluded from further analyses by checking the "Exclude" box.
Scatter plot
PCA-Plot
of 11 Affymetrix data files
HCLUST-Plot
of 11 Affymetrix data files
Previous
Next
Step 3 of 4
idle
Manual

Evaluating QC results

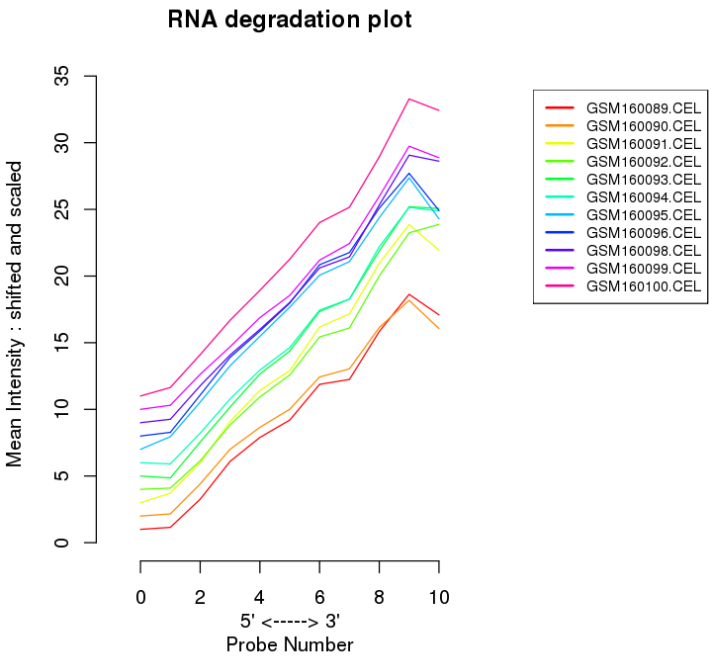
One plot of each kind is reproduced below and shows that RobiNA generates classical QC plots showing how 'good' the data is and how well it divides according to the defined groups.

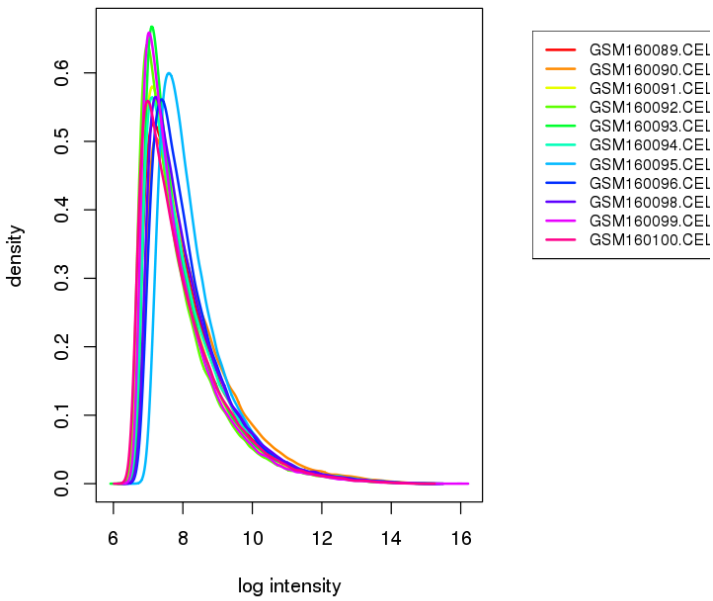
Details for each QC-Plot type	[Collapse]

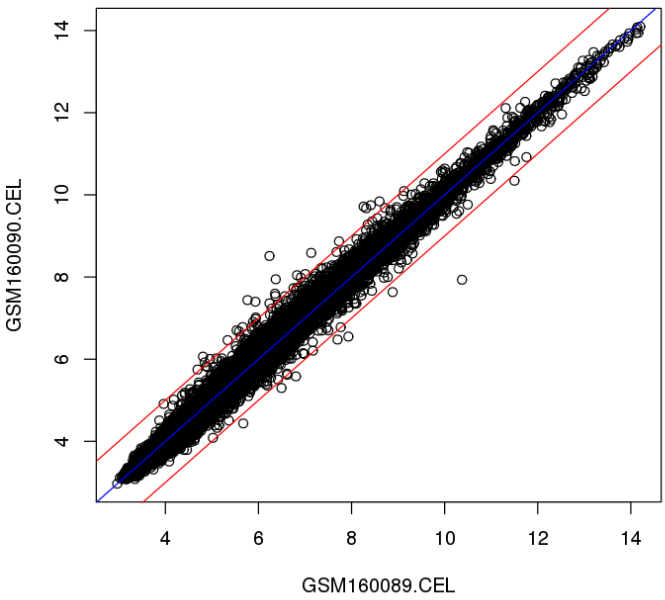




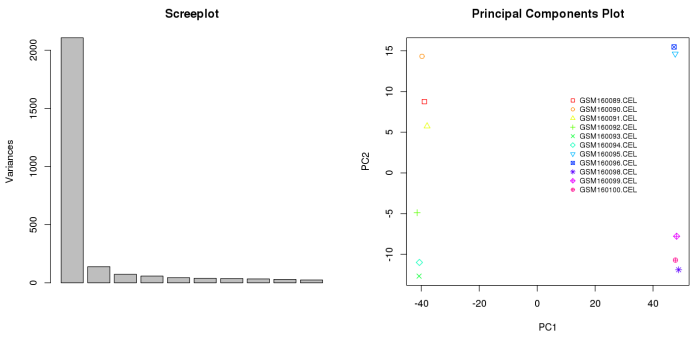
one for each sample



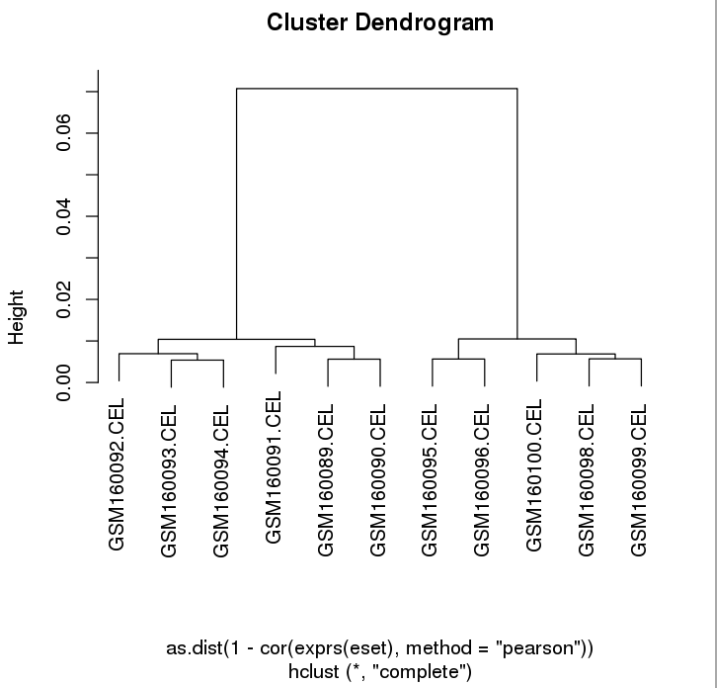




<- one for each pairwise sample comparison (n^2)

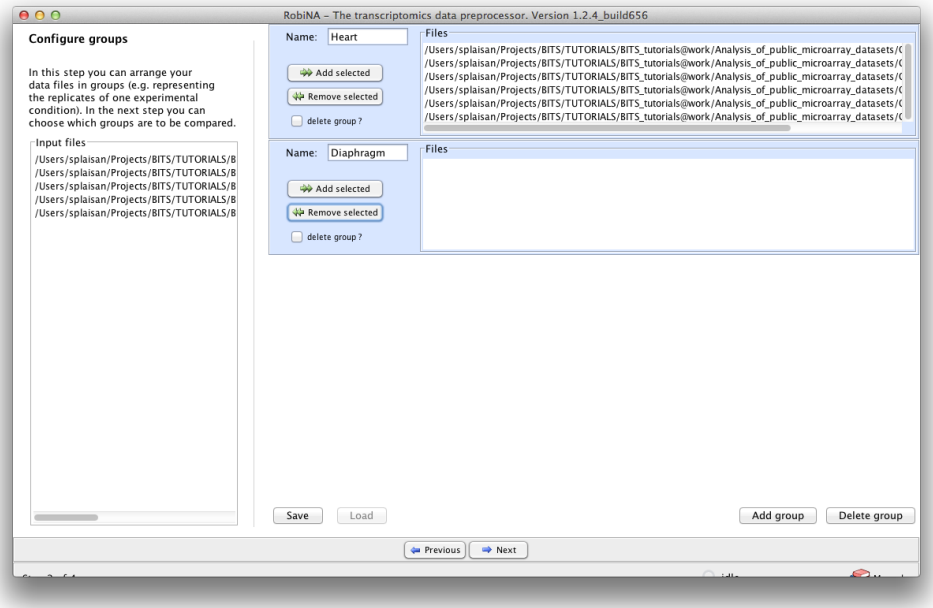


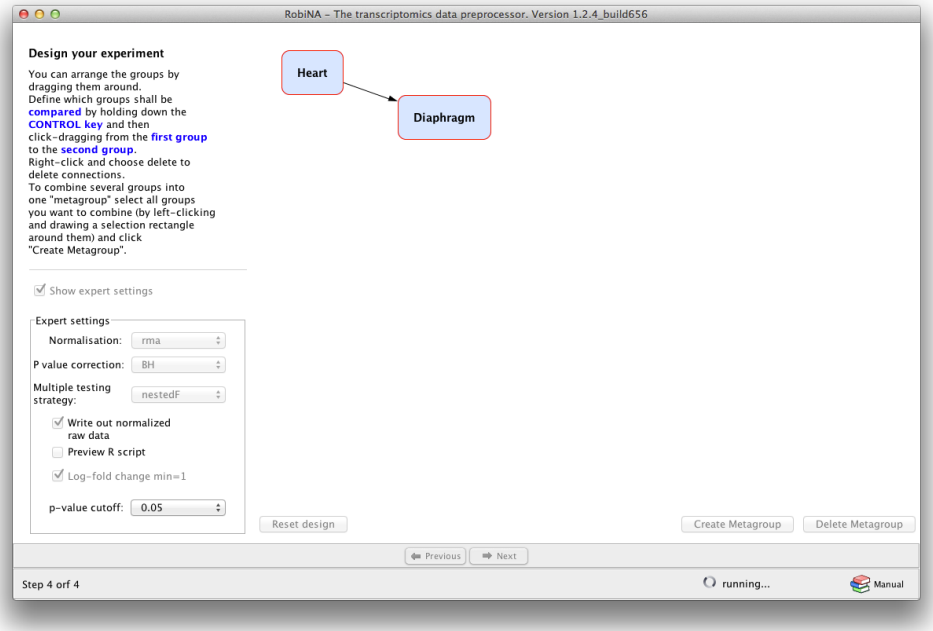
distribution of the variance across principal components and PCA plot



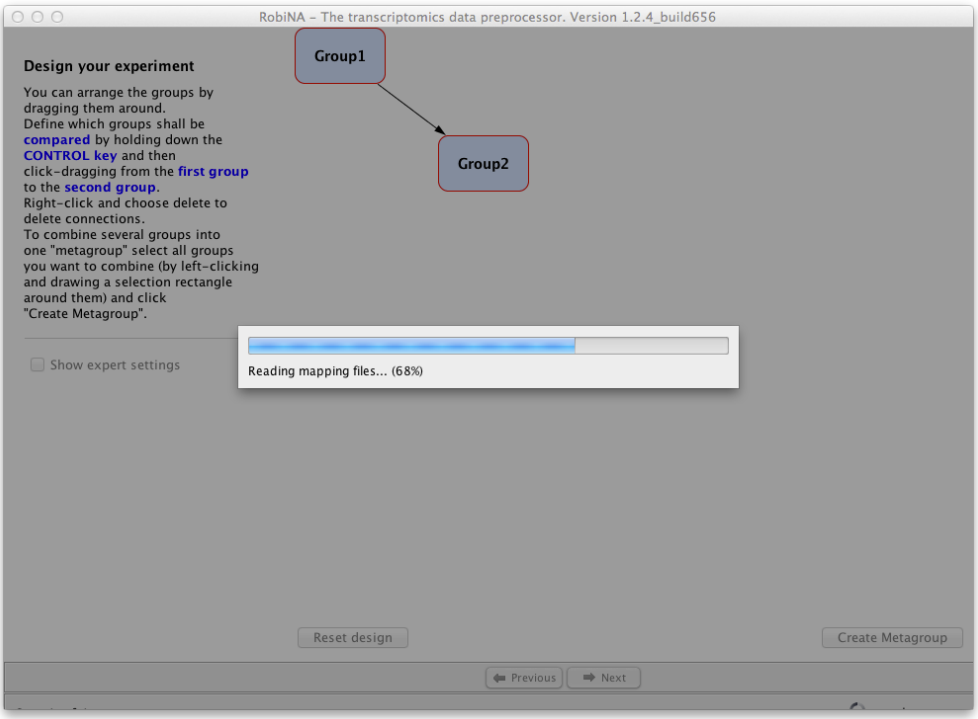
Define design

In this part, we define contrasts and start the differential analysis. Please note that complex design are possible here by defining 'metagroups' next to the sample groups (eg. mouse background, stimulant, ...). This is not the aim of this very simplistic design comparing two tissues. Please refer to the software documentation for more detailed examples.

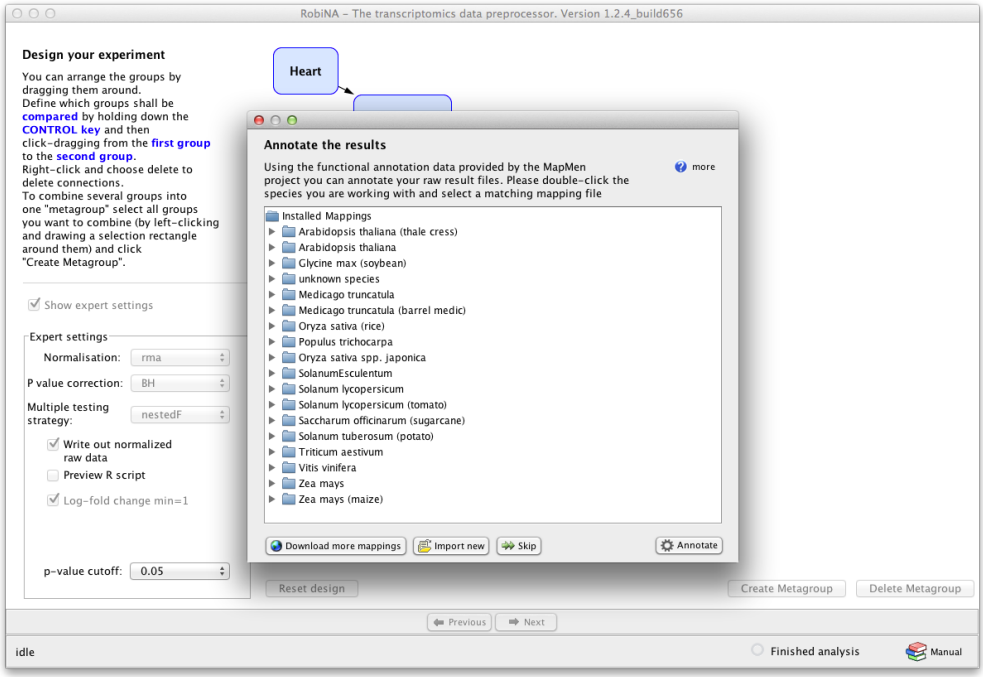



RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Design your experiment
You can arrange the groups by dragging them around.
Define which groups shall be compared by holding down the CONTROL key and then click-dragging from the first group to the second group.
Right-click and choose delete to delete connections.
To combine several groups into one "metagroup" select all groups you want to combine (by left-clicking and drawing a selection rectangle around them) and click "Create Metagroup".
Heart
Diaphragm
Show expert settings
Expert settings
Normalisation: rma
P value correction: BH
Multiple testing strategy: nestedF
Write out normalized raw data
Preview R script
Log-fold change min=1
p-value cutoff: 0.05
Reset design
Create Metagroup
Delete Metagroup
Previous
Next
Step 4 orf 4
running...
Manual

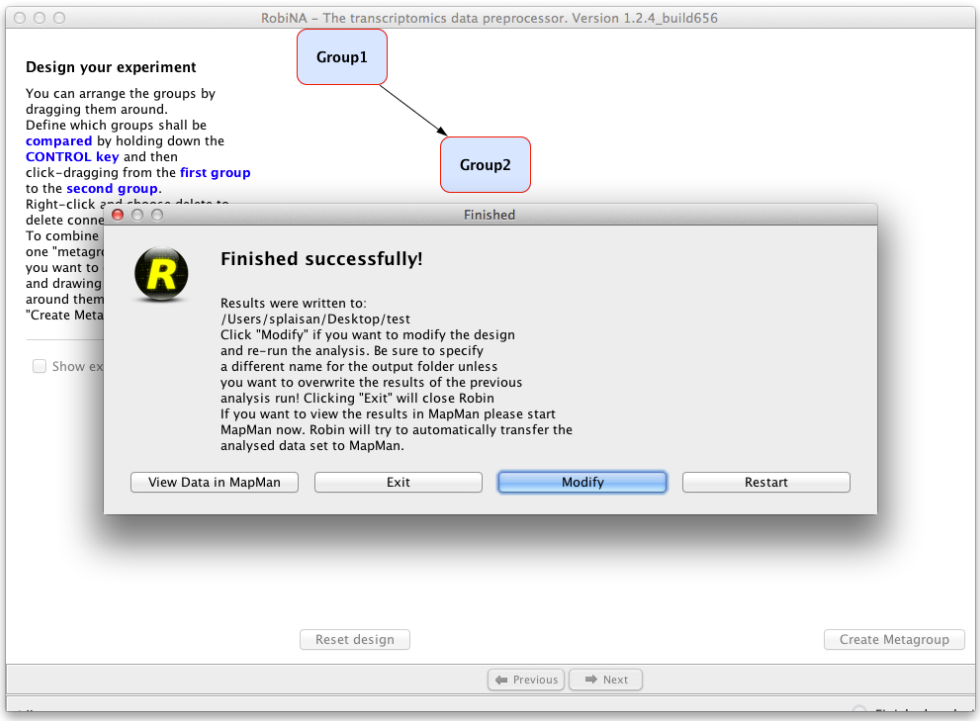
Computing differential expression


RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Design your experiment
You can arrange the groups by dragging them around.
Define which groups shall be compared by holding down the CONTROL key and then click-dragging from the first group to the second group.
Right-click and choose delete to delete connections.
To combine several groups into one "metagroup" select all groups you want to combine (by left-clicking and drawing a selection rectangle around them) and click "Create Metagroup".
Group1
Group2
Show expert settings
Reading mapping files... (68%)
Reset design
Create Metagroup
Previous
Next

RobiNA was developed for plants and can use annotation files from the website to add gene descriptions to the differential expression table. For non-plant data, users will need to add these annotations using external tools


RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Design your experiment
You can arrange the groups by dragging them around.
Define which groups shall be compared by holding down the CONTROL key and then click-dragging from the first group to the second group.
Right-click and choose delete to delete connections.
To combine several groups into one "metagroup" select all groups you want to combine (by left-clicking and drawing a selection rectangle around them) and click "Create Metagroup".
Heart
Annotate the results
Using the functional annotation data provided by the MapMen project you can annotate your raw result files. Please double-click the species you are working with and select a matching mapping file
more
Installed Mappings
Arabidopsis thaliana (thale cress)
Arabidopsis thaliana
Glycine max (soybean)
unknown species
Medicago truncatula
Medicago truncatula (barrel medic)
Oryza sativa (rice)
Populus trichocarpa
Oryza sativa spp. japonica
SolanumEsculentum
Solanum lycopersicum
Solanum lycopersicum (tomato)
Saccharum officinarum (sugarcane)
Solanum tuberosum (potato)
Triticum aestivum
Vitis vinifera
Zea mays
Zea mays (maize)
Download more mappings
Import new
Skip
Annotate
Show expert settings
Expert settings
Normalisation: rma
P value correction: BH
Multiple testing strategy: nestedF
Write out normalized raw data
Preview R script
Log-fold change min=1
p-value cutoff: 0.05
Reset design
Create Metagroup
Delete Metagroup
Previous
Next
idle
Finished analysis
Manual

We choose 'Skip' as we do not have annotation files for rat. The differential expression gets computed. When done, we select 'Exit' from the front window.

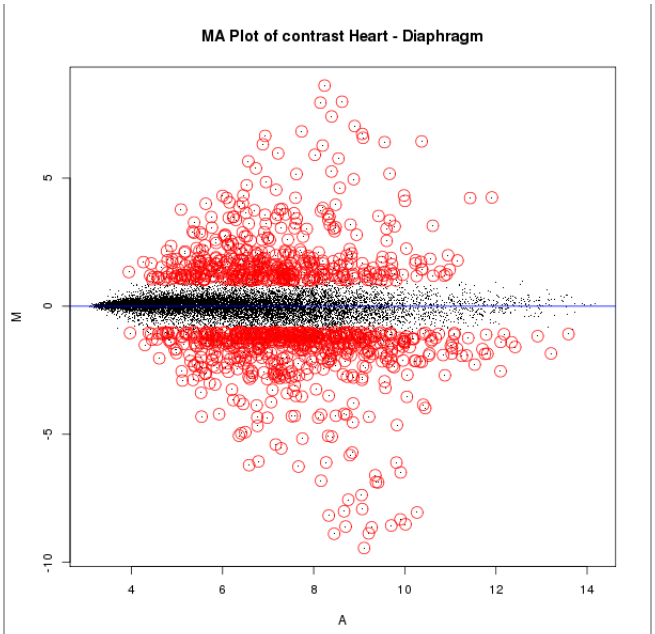

RobiNA – The transcriptomics data preprocessor. Version 1.2.4_build656
Design your experiment
You can arrange the groups by dragging them around.
Define which groups shall be compared by holding down the CONTROL key and then click-dragging from the first group to the second group.
Group1
Group2
Finished
Finished successfully!
Results were written to:
/Users/splaisan/Desktop/test
Click "Modify" if you want to modify the design and re-run the analysis. Be sure to specify a different name for the output folder unless you want to overwrite the results of the previous analysis run! Clicking "Exit" will close Robin
If you want to view the results in MapMan please start MapMan now. Robin will try to automatically transfer the analysed data set to MapMan.
View Data in MapMan
Exit
Modify
Restart
Reset design
Create Metagroup
Previous
Next

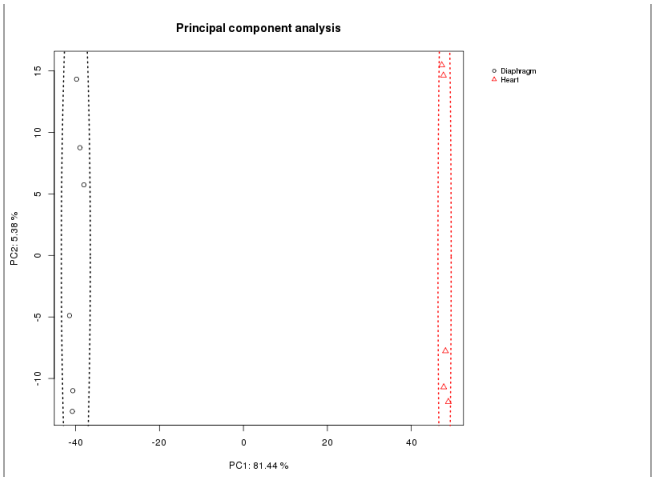
Reviewing DE results

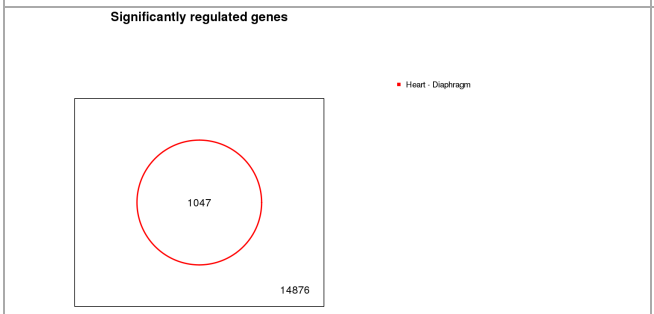
RobiNA saves a number of files to the disk in its folder structure. Text tables can be found in the 'detailed results' folder and are partly reproduced below. A short summary lists the parameters applied to perform the analysis while two tables report the **full results** and the **top-100 results** after sorting lines in **decreasing adj.P.Val** order. Three PNG pictures (below) report the Up-regulated gene count, DE-gene count and Down-regulated gene-count respectively.

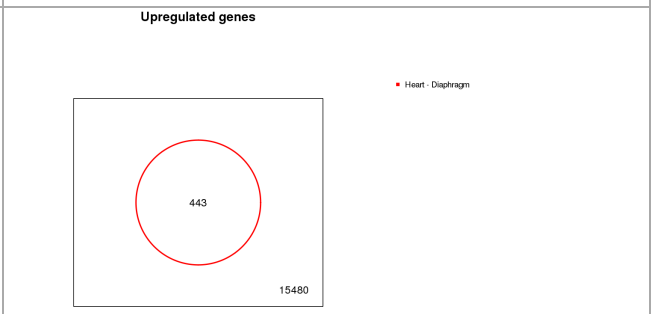
Final plots summarize key facets of the analysis as done in classical MA analysis. The left plot shows in red probesets that are differentially expressed with the selected statistics; the right plot confirms that the samples segregate as defined in the grouping. The bottom simple venn diagrams report counts of DE probesets (Total, Up, Down).

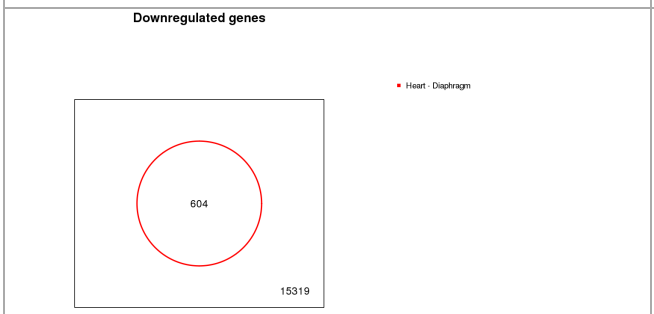
Plots	[Collapse]











In addition to the plots, several text tables are saved in 'detailed_results' and sampled below.

Analysis summary *[Collapse]*

```
#################################
# Robin affymetrix data analysis summary
# RobiNA-results
#  8/18/2014; 16:35:28
#################################

# Input files:
```

```
# Normalization settings for quality control
normalization method: rma
P-value correction method: BH
analysis strategy: Limma

# Normalization settings for main analysis
normalization method: rma
P-value correction method: BH
Multiple testing strategy: nestedF
P-value cut-off value for significant differential expression: 0.05
Genes that showed a log2-fold change smaller than two ignored: yes

# The analysis produced the following warnings
none
```

Head of 'full_table_Heart - Diaphragm.txt' *[Collapse]*

```
ID          logFC              AveExpr           t                 P.Value               adj.P.Val             B
1371247_at  8.8690486752098    9.2066061992338   138.180384786338  7.72673860255015e-23  1.23032858768406e-18  40.3028038935592
1370412_at  8.00700839973324   8.66636177255196  112.671243745562  1.24362501567467e-21  9.90112056229389e-18  38.5147635203019
1387787_at  8.51384348690296   10.0136513602682  107.791218026326  2.27212516655303e-21  1.2059683009008e-17   38.0893208171574
1372195_at  9.44110458586408   9.1061642897453   100.528072562373  5.87151716284703e-21  2.01180346646277e-17  37.3935367235781
1370971_at  8.88483556476391   8.44782914943697  99.9889507276205  6.31728778013807e-21  2.01180346646277e-17  37.3386444895643
1367896_at  8.17319442370653   8.32878048386055  97.7761545656647  8.56617301567765e-21  2.27331954881059e-17  37.1083132290366
1374248_at  7.92028193041167   9.06490481778715  95.9358621573261  1.10936083087999e-20  2.52347893001459e-17  36.9103955684498
1368093_at  -8.61265104186292  8.23846334693981  -81.9875192287634 9.40341381711377e-20  1.87163197762378e-16  35.195893580546
1388139_at  8.60278366808529   8.70065810953839  80.0047250421511  1.31181945592154e-19  2.32090013295985e-16  34.9170369337463
```

Head of 'mean_rma_normalized_expression_values.txt' *[Collapse]*

```
Identifier  Heart              Diaphragm
1367452_at  9.5293566110383    9.5889443412077
1367453_at  10.1758428259477   10.1726921462891
1367454_at  8.87231561088045   8.86593395245015
1367455_at  10.4751392501094   10.3325900097593
1367456_at  10.7698324122072   10.286304116399
1367457_at  8.75165515455833   8.73159768841635
1367458_at  7.54000953569876   7.71781966096578
1367459_at  11.0420305904582   11.1519497142209
1367460_at  9.90893064678104   9.5235636384544
```

Head of 'raw_rma_normalized_expression_values.txt' *[Collapse]*

```
Identifier  GSM160089.CEL     GSM160090.CEL     GSM160091.CEL     GSM160092.CEL     GSM160093.CEL     GSM160094.CEL     GSM160095.CEL     GSM160096.CEL     GS
1367452_at  9.64295883656058  9.65178912787566  9.32863584112902  9.48475646698994  9.4716451160078   9.59635427766681  9.65872553515732  9.63503499785589  9.
1367453_at  10.1553644909771  10.2499020063438  10.1861859768652  10.0965629262007  10.159434523851   10.2076070314483  10.2269807954293  10.195253118062   10
1367454_at  8.73905780465407  9.08252980064849  9.07161018095027  8.68762900735943  8.81499836330383  8.83806850836661  8.88654295440762  9.05310137910709  8.
1367455_at  10.4763615418993  10.4471239467834  10.5710504733759  10.3264804870339  10.4745349426514  10.5552841089126  10.5959387062089  10.5139881779964  10
1367456_at  10.9984593902943  10.8441477983143  10.673570257498   10.6700249981427  10.6750331606405  10.7577588683534  10.1986166073672  10.3974448991471  10
1367457_at  8.78904682538038  8.72229122394877  8.67050724938203  8.9448960477874   8.72229122394877  8.66089835690261  8.76340465790531  8.76513510798621  8.
1367458_at  7.53098410601309  7.35718220143744  7.71709223686958  7.65558449687549  7.45098048336888  7.52823368962808  7.7461437721208   7.40572408979464  7.
1367459_at  11.1105760949228  11.2159762048115  10.9101244239086  11.149432063699   10.8953374468102  10.970737308597   11.2355512291063  11.2441035772632  11
1367460_at  9.96660388228079  9.84511446842093  9.94477628710098  9.91078389672863  9.84447947912025  9.94182586703467  9.87080073408934  9.70116063385167  9.
```

Adding annotations to the RobiNA data

Because RobiNA does not support non-plant organisms, it saves the data in a quite anonymous way with only probe IDs. The code below is borrowed from the GEO2R script and adds gene symbols and additional annotations to the RobiNA table.

R-code to annotate the RobiNA data *[Collapse]*

```
# Add annotations to the RobiNA full table
# the code below is borrowed to the GEO2R code and adapted
library(GEOquery)

# make sure that the surrent directory is set to folder enclosing the 'RobiNA-results' folder
base <- getwd()

# load RobiNA full table in a data-frame
robina.folder <- paste(base, "RobiNA-results", "detailed_results", sep="/")
robina.data <- paste(robina.folder, "full_table_Heart - Diaphragm.txt", sep="/")
robina.full <- read.delim(robina.data, as.is = TRUE)

# load NCBI platform annotation
gpl <- "GPL341"
platf <- getGEO(gpl, AnnotGPL=TRUE, destdir = base)
ncbifd <- data.frame(attr(dataTable(platf), "table"))

# replace original platform annotation
data <- merge(robina.full, ncbifd, by="ID")
data <- data[order(data$P.Value), ]  # restore correct order

# preview first 10 columns
head(data)[,c(1:10)]
```

```
> head(data)[,c(1:10)]
              ID    logFC   AveExpr         t      P.Value    adj.P.Val        B
3796  1371247_at 8.869049  9.206606 138.18038 7.726739e-23 1.230329e-18 40.30280
2961  1370412_at 8.007008  8.666362 112.67124 1.243625e-21 9.901121e-18 38.51476
11906 1387787_at 8.513843 10.013651 107.79122 2.272125e-21 1.205968e-17 38.08932
4744  1372195_at 9.441105  9.106164 100.52807 5.871517e-21 2.011803e-17 37.39354
3520  1370971_at 8.884836  8.447829  99.98895 6.317288e-21 2.011803e-17 37.33864
445   1367896_at 8.173194  8.328780  97.77615 8.566173e-21 2.273320e-17 37.10831
                                                                                Gene.title Gene.symbol
3796                                                          troponin T type 3 (skeletal, fast)       Tnnt3
2961                                                          troponin T type 1 (skeletal, slow)       Tnnt1
11906                                        myosin light chain, phosphorylatable, fast skeletal muscle       Mylpf
4744                                                            troponin C type 2 (fast)       Tnnc2
3520  myosin, heavy chain 2, skeletal muscle, adult///myosin, heavy chain 1, skeletal muscle, adult Myh2///Myh1
445                                                                carbonic anhydrase 3        Car3
             Gene.ID
3796           24838
```

```
2961           171409
11906           24584
4744           296369
3520  691644///287408
445             54232
```

```
# save to new file
outfile <- paste(robina.folder, "GSE6943_DE-Robina.txt", sep="/")
write.table(data, file=outfile, row.names=F, sep="\t", quote=FALSE)

colnames(data)
```

The resulting file has gained 20 columns in addition to the first 7 created by RobiNA

```
 [1] "ID"                     "logFC"                  "AveExpr"                "t"
 [5] "P.Value"                "adj.P.Val"              "B"                      "Gene.title"
 [9] "Gene.symbol"            "Gene.ID"                "UniGene.title"          "UniGene.symbol"
[13] "UniGene.ID"             "Nucleotide.Title"       "GI"                     "GenBank.Accession"
[17] "Platform_CLONEID"       "Platform_ORF"           "Platform_SPOTID"        "Chromosome.location"
[21] "Chromosome.annotation" "GO.Function"            "GO.Process"             "GO.Component"
[25] "GO.Function.ID"         "GO.Process.ID"          "GO.Component.ID"
```

Conclusion

RobiNA is a wrapper for R-code, developed and standardized by the authors to run reproducibly and do as they are expected. Although simple in layout, RobiNA appears as a quite performant alternative top more top-notch analysis methods reported in the literature. A computer with sufficient resources is wished to perform QC steps in a reasonable time but current strong laptops and desktops should do the job.

For those who cannot afford the more expensive CLC-Main workbench and do not wish to learn [R] and Bioconductor, RobiNA seems a good choice when in need of MA analysis (and it can also do standard RNASeq analysis - as will be described in a separate hands-on)

The complete file is accessible **here** (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex4-files/RobiNA-results/detailed_results/GSE6943_DE-Robina.txt) (use right-click download linked file as or navigate from the page bottom-link)

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑
   Marc Lohse, Adriano Nunes-Nesi, Peter Krüger, Axel Nagel, Jan Hannemann, Federico M Giorgi, Liam Childs, Sonia Osorio, Dirk Walther, Joachim Selbig, Nese Sreenivasulu, Mark Stitt, Alisdair R Fernie, Björn Usadel
   Robin: an intuitive wizard application for R-based expression microarray quality assessment and analysis.
   Plant Physiol.: 2010, 153(2);642-51
   [PubMed:20388663] ##WORLDCAT## [DOI] (I p)

   ```
   http://mapman.gabipd.org/web/guest/robin
   ```

2. ↑ [2.0] [2.1] http://www.affymetrix.com/estore/

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=PubMA_Exercise.4&oldid=11748"
Category: PUBMA2014

- This page was last modified on 16 October 2014, at 09:49.
- This page has been accessed 181 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

PubMA Exercise.5

From BioWareWIKI

Functional enrichment of the obtained lists to identify key biological functions

---

Contents

- 1 Why we are not yet done!
- 2 **Functional enrichment Analysis** of the **RobiNA** DE-data
  - 2.1 Preparing probe lists for enrichment testing
  - 2.2 **DAVID**: father of enrichment tools
  - 2.3 Current web-based Enrichment-tools
    - 2.3.1 **Enrich**
    - 2.3.2 **Webgestalt**
- 3 Conclusion
- 4 download exercise files

Why we are not yet done!

Once upon a time, scientists worked for a lifetime on one or few genes, they read all what came published about their favorite proteins and did not need any computer to help them follow up and understand published data.

This time now belongs to the past and modern biologists need to cope with publication frequency far higher than their reading speed. happy or not, they have to rely on computers for some of the tasks they used to do manually.

Analysis of MA data, similarly to any other high throughput technology, generates thousands of lines of results out of which hundreds are statistically significant. It is therefore very unlikely that the biologist evaluate each line and identify the proteins (genes products) that are significantly altered in expression and may be responsible for the biology under investigation.

The approach consisting in recognizing genes in the list and selecting them for validation may seem appropriate but will unlikely lead to any discoveries. As the main need for publication is to find novelty, this method is pretty much useless.

A better way to analyze and prioritize targets from a screen is to consider the biological functions and pathways that include [are-enriched-in] differentially expressed genes. This can be done after adding ontology annotations to the data and using these added column to identify 'functions', 'diseases', 'pathways' or any ontologies terms that are enriched in the DE-set et as compared to the full set of genes measured by the platform. Again, this apparently straightforward statistical testing can be quite lengthy if you consider hundreds of available ontologies and hundreds to thousand genes to annotate.

Hypergeometric T-test and **Gene set enrichment analysis** (GSEA) are the two mainly used statistical approaches to identify enrichment based on gene lists. A number of standalone and Web tools implement these methods and is falls beyond the scope of this training to list them all or to argue for one or another. We instead present a few alternative tools that will accept the data obtained in the former exercises and process it to find enriched ontology terms.

Functional enrichment Analysis of the RobiNA DE-data

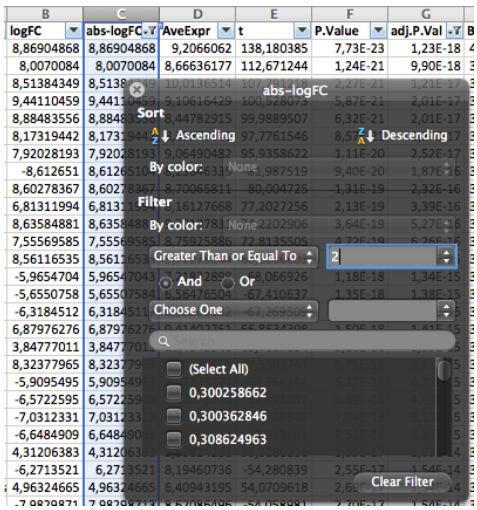
In order to continue with the most complete and easy workflow, we use here the **RobiNA** table obtained by de-novo analysis of 11 of the 12 samples (one CEL file being damaged on the GEO repository)

Preparing probe lists for enrichment testing

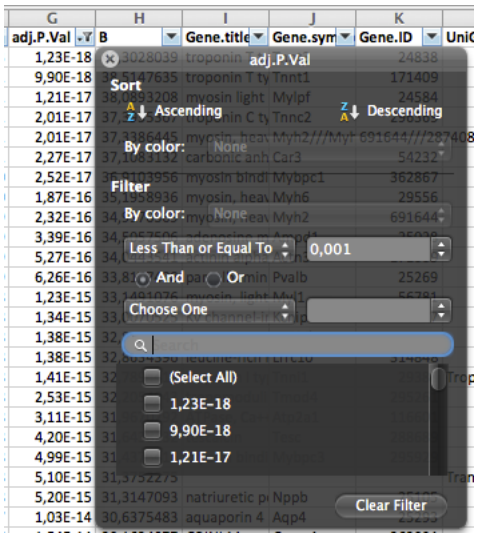
Web tools will require probe or gene lists to compute enrichment, they will not take into account the degree of DE or the confidence in that DE both of which are left to the user to filter.

We can produce these two lists using Excel (better would be R) in few easy steps

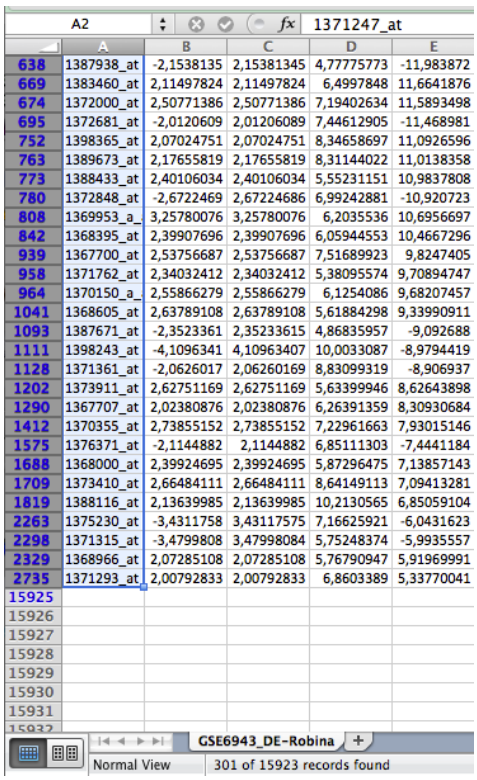
- import the table in excel, taking care of protecting gene symbols against interpretation
- add a column with absolute value of logFC
- filter on the abs(logFC) with a minimal cutoff of 2 (four fold DE)



- filter on the 'adj.P.Val' column with a maximum of 0.001



The list of 15924 probes is reduced to 301 by this double filtering



	A	B	C	D	E
638	1387938_at	-2,1538135	2,15381345	4,77775773	-11,983872
669	1383460_at	2,11497824	2,11497824	6,4997848	11,6641876
674	1372000_at	2,50771386	2,50771386	7,19402634	11,5893498
695	1372681_at	-2,0120609	2,01206089	7,44612905	-11,468981
752	1398365_at	2,07024751	2,07024751	8,34658697	11,0926596
763	1389673_at	2,17655819	2,17655819	8,31144022	11,0138358
773	1388433_at	2,40106034	2,40106034	5,55231151	10,9837808
780	1372848_at	-2,6722469	2,67224686	6,99242881	-10,920723
808	1369953_a_	3,25780076	3,25780076	6,2035536	10,6956697
842	1368395_at	2,39907696	2,39907696	6,05944553	10,4667296
939	1367700_at	2,53756687	2,53756687	7,51689923	9,8247405
958	1371762_at	2,34032412	2,34032412	5,38095574	9,70894747
964	1370150_a_	2,55866279	2,55866279	6,1254086	9,68207457
1041	1368605_at	2,63789108	2,63789108	5,61884298	9,33990911
1093	1387671_at	-2,3523361	2,35233615	4,86835957	-9,092688
1111	1398243_at	-4,1096341	4,10963407	10,0033087	-8,9794419
1128	1371361_at	-2,0626017	2,06260169	8,83099319	-8,906937
1202	1373911_at	2,62751169	2,62751169	5,63399946	8,62643898
1290	1367707_at	2,02380876	2,02380876	6,26391359	8,30930684
1412	1370355_at	2,73855152	2,73855152	7,22961663	7,93015146
1575	1376371_at	-2,1144882	2,1144882	6,85111303	-7,4441184
1688	1368000_at	2,39924695	2,39924695	5,87296475	7,13857143
1709	1373410_at	2,66484111	2,66484111	8,64149113	7,09413281
1819	1388116_at	2,13639985	2,13639985	10,2130565	6,85059104
2263	1375230_at	-3,4311758	3,43117575	7,16625921	-6,0431623
2298	1371315_at	-3,4799808	3,47998084	5,75248374	-5,9935557
2329	1368966_at	2,07285108	2,07285108	5,76790947	5,91969991
2735	1371293_at	2,00792833	2,00792833	6,8603389	5,33770041
15925					
15926					
15927					
15928					
15929					
15930					
15931					
15932					

- export probes and gene lists to text files
  - **301 DE probes** (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex5-files/RobiNA-DE-probes_LFC2-FDR0.001.txt)
  - **238 DE genes** (unique) (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex5-files/RobiNA-DE-genes_LFC2-FDR0.001.txt)
  - all probes in the rat array (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex5-files/RobiNA-all-probes.txt)

DAVID: father of enrichment tools

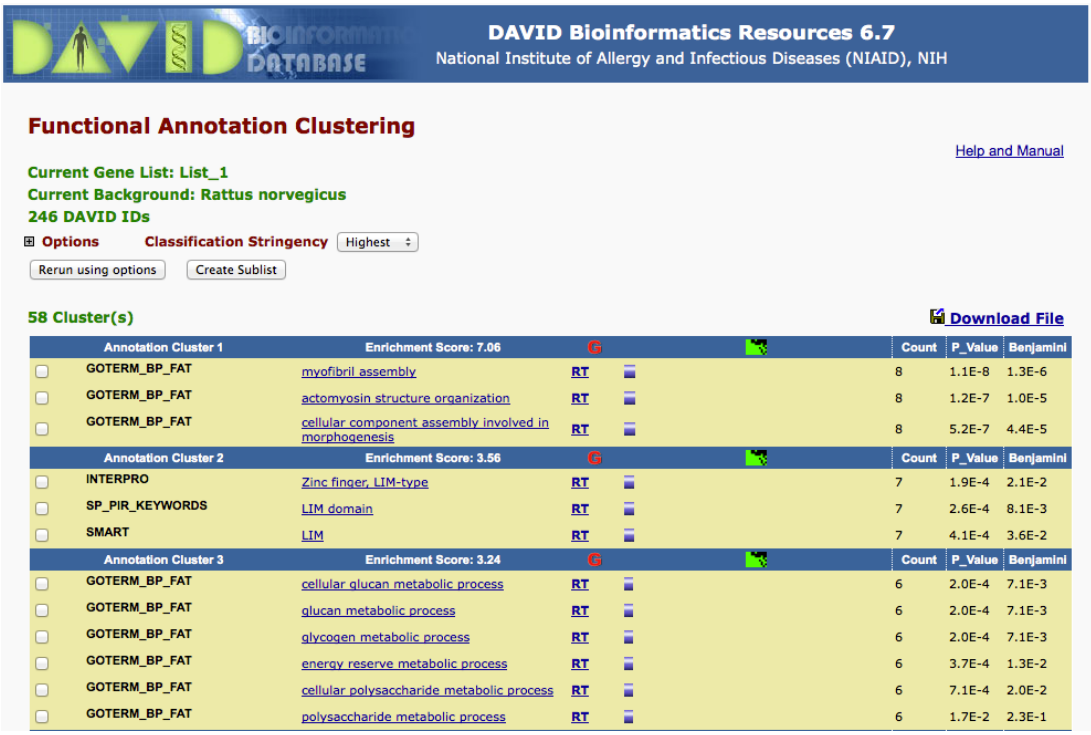
The canonical web tool is **DAVID**[1]. **DAVID** has been around since 1997 and is stil very popular although its interface is quite outdated. A recent nature Protocol paper[2] will help you start using DAVID.

In order to use DAVID, we need to spit our data in two groups:

- genes (probe IDs) considered differentially expressed (TEST)
- the remaining of the genes (probes) present on the platform (BACKGROUND) - note that the DAVID 'Background' tab allows selecting the actual MA chip (Rat Genome RAE230A Array)

This division is very important to obtain good results and ensures that only functions represented by genes assessed in the experiment will be considered, although this is normally true with modern platform where almost all known genes are present

- When the full Rat genome is taken as background, we obtain relatively high confidence predictions



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Rattus norvegicus
246 DAVID IDs

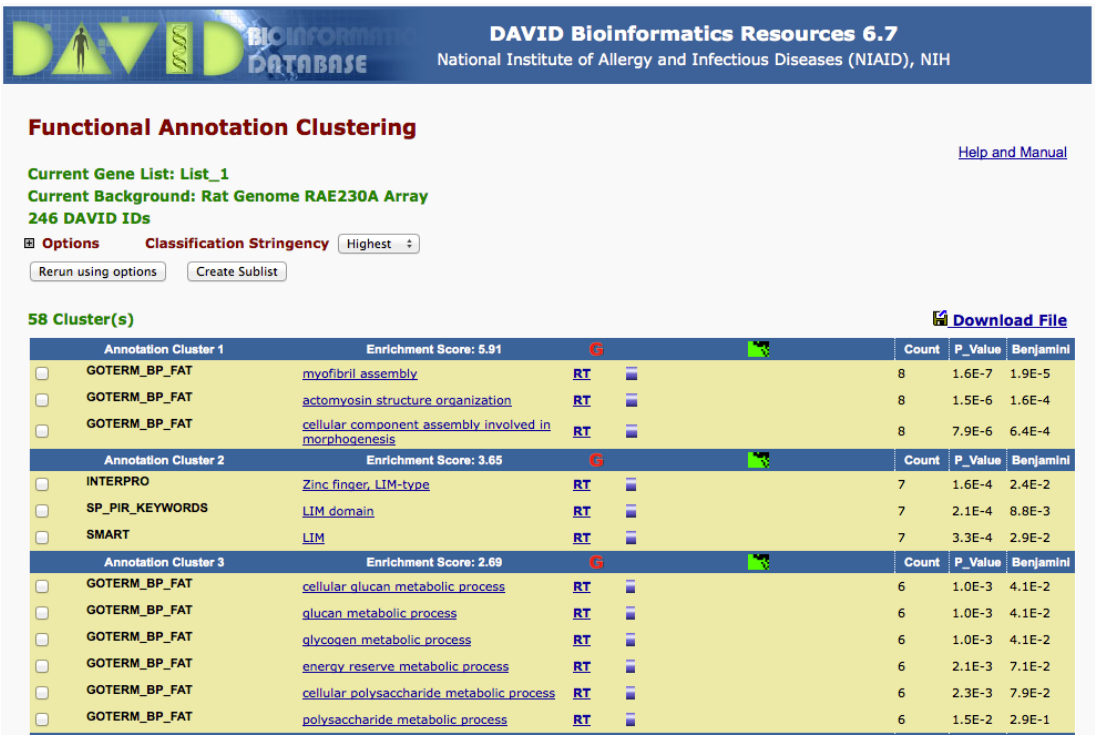
Options Classification Stringency: Highest

Rerun using options Create Sublist

58 Cluster(s) Download File

Annotation Cluster 1	**Enrichment Score: 7.06**				**Count**	**P_Value**	**Benjamini**
GOTERM_BP_FAT	myofibril assembly	RT			8	1.1E-8	1.3E-6
GOTERM_BP_FAT	actomyosin structure organization	RT			8	1.2E-7	1.0E-5
GOTERM_BP_FAT	cellular component assembly involved in morphogenesis	RT			8	5.2E-7	4.4E-5
Annotation Cluster 2	**Enrichment Score: 3.56**				**Count**	**P_Value**	**Benjamini**
INTERPRO	Zinc finger, LIM-type	RT			7	1.9E-4	2.1E-2
SP_PIR_KEYWORDS	LIM domain	RT			7	2.6E-4	8.1E-3
SMART	LIM	RT			7	4.1E-4	3.6E-2
Annotation Cluster 3	**Enrichment Score: 3.24**				**Count**	**P_Value**	**Benjamini**
GOTERM_BP_FAT	cellular glucan metabolic process	RT			6	2.0E-4	7.1E-3
GOTERM_BP_FAT	glucan metabolic process	RT			6	2.0E-4	7.1E-3
GOTERM_BP_FAT	glycogen metabolic process	RT			6	2.0E-4	7.1E-3
GOTERM_BP_FAT	energy reserve metabolic process	RT			6	3.7E-4	1.3E-2
GOTERM_BP_FAT	cellular polysaccharide metabolic process	RT			6	7.1E-4	2.0E-2
GOTERM_BP_FAT	polysaccharide metabolic process	RT			6	1.7E-2	2.3E-1

- By contrast; when the true background is set to what the RAE230A really covers, a lower confidence is obtained. This is not a major issue here but when reporting p.values, you should always be careful to use the correct background in order not to overestimate your findings.



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Rat Genome RAE230A Array
246 DAVID IDs

Options Classification Stringency: Highest

Rerun using options Create Sublist

58 Cluster(s) Download File

Annotation Cluster 1	**Enrichment Score: 5.91**				**Count**	**P_Value**	**Benjamini**
GOTERM_BP_FAT	myofibril assembly	RT			8	1.6E-7	1.9E-5
GOTERM_BP_FAT	actomyosin structure organization	RT			8	1.5E-6	1.6E-4
GOTERM_BP_FAT	cellular component assembly involved in morphogenesis	RT			8	7.9E-6	6.4E-4
Annotation Cluster 2	**Enrichment Score: 3.65**				**Count**	**P_Value**	**Benjamini**
INTERPRO	Zinc finger, LIM-type	RT			7	1.6E-4	2.4E-2
SP_PIR_KEYWORDS	LIM domain	RT			7	2.1E-4	8.8E-3
SMART	LIM	RT			7	3.3E-4	2.9E-2
Annotation Cluster 3	**Enrichment Score: 2.69**				**Count**	**P_Value**	**Benjamini**
GOTERM_BP_FAT	cellular glucan metabolic process	RT			6	1.0E-3	4.1E-2
GOTERM_BP_FAT	glucan metabolic process	RT			6	1.0E-3	4.1E-2
GOTERM_BP_FAT	glycogen metabolic process	RT			6	1.0E-3	4.1E-2
GOTERM_BP_FAT	energy reserve metabolic process	RT			6	2.1E-3	7.1E-2
GOTERM_BP_FAT	cellular polysaccharide metabolic process	RT			6	2.3E-3	7.9E-2
GOTERM_BP_FAT	polysaccharide metabolic process	RT			6	1.5E-2	2.9E-1

A typical DAVID plot for the top cluster

myosin, light polypeptide 1
myosin, heavy polypeptide 2, skeletal muscle, adult; myosin, heavy polypeptide 1, skeletal muscle, adult
leucine-rich repeat-containing 10
myosin light chain, phosphorylatable, fast skeletal muscle
myosin, heavy chain 4, skeletal muscle
myosin, heavy chain 6, cardiac muscle, alpha
synemin, intermediate filament protein
troponin I type 1 (skeletal, slow)
troponin T type 1 (skeletal, slow)
troponin C type 2 (fast)
troponin I type 3 (cardiac)
myosin binding protein C, cardiac
troponin T type 3 (skeletal, fast)
troponin I type 2 (skeletal, fast)
troponin T type 2 (cardiac)
myosin, light chain 7, regulatory
ankyrin repeat domain 1 (cardiac muscle)
calsequestrin 1 (fast-twitch, skeletal muscle)
ATPase, Ca++ transporting, cardiac muscle, fast twitch 1
ankyrin repeat domain 23
actin, alpha, cardiac muscle 1; smooth muscle alpha-actin
titin
ryanodine receptor 2, cardiac
nebulin
junctophilin 1
keratin 19
PDZ and LIM domain 3
cysteine and glycine-rich protein 3
four and a half LIM domains 3
calcium channel, voltage-dependent, L type, alpha 1C subunit
four and a half LIM domains 2
GO 0043292~contractile fiber
GO 0030016~myofibril
GO 0044449~contractile fiber part
GO 0030017~sarcomere
GO 0031674~I band
GO 0030018~Z disc

Current web-based Enrichment-tools

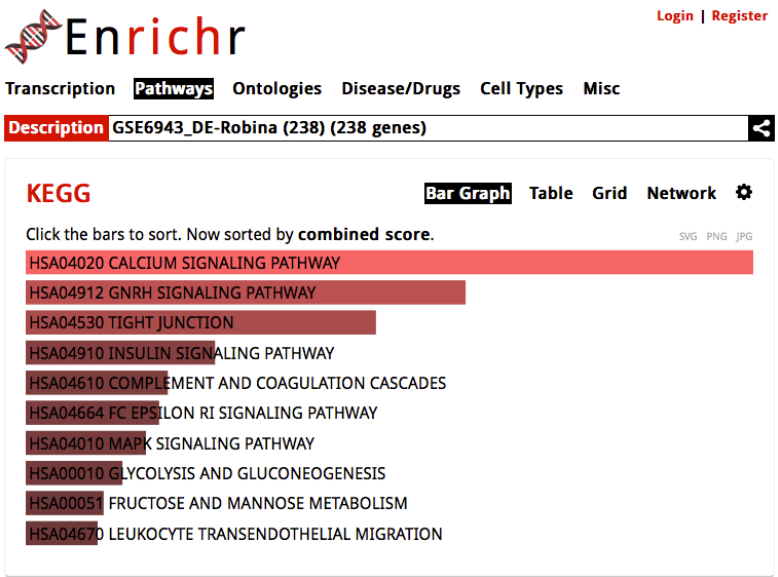
We chose to show you only two recent tools in this section. Many others are available and you are welcome to try them and share your experience back with us.

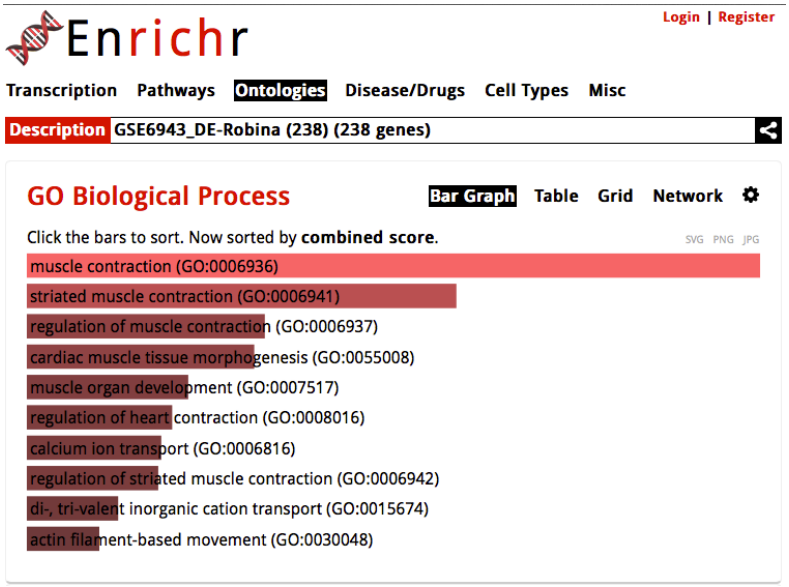
Enrich

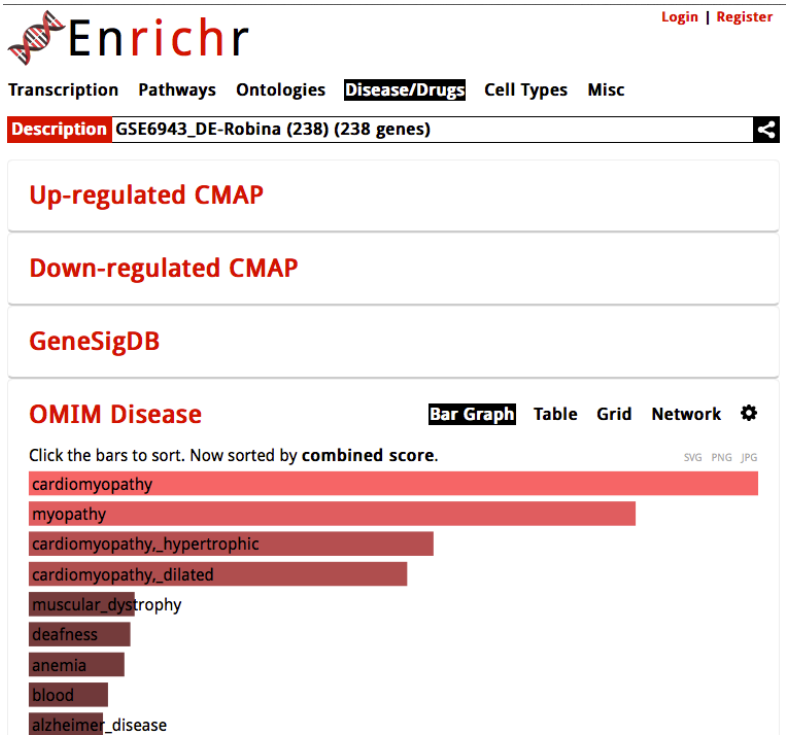
Enrichr [3] and its associated tool **Lists2Networks** [4] use a respectable number of sources to compute enrichment (source-list: http://amp.pharm.mssm.edu/Enrichr/index.html#stats). To learn more about Enrich, please refer to their **FAQ page (http://amp.pharm.mssm.edu/Enrichr/#help)**.

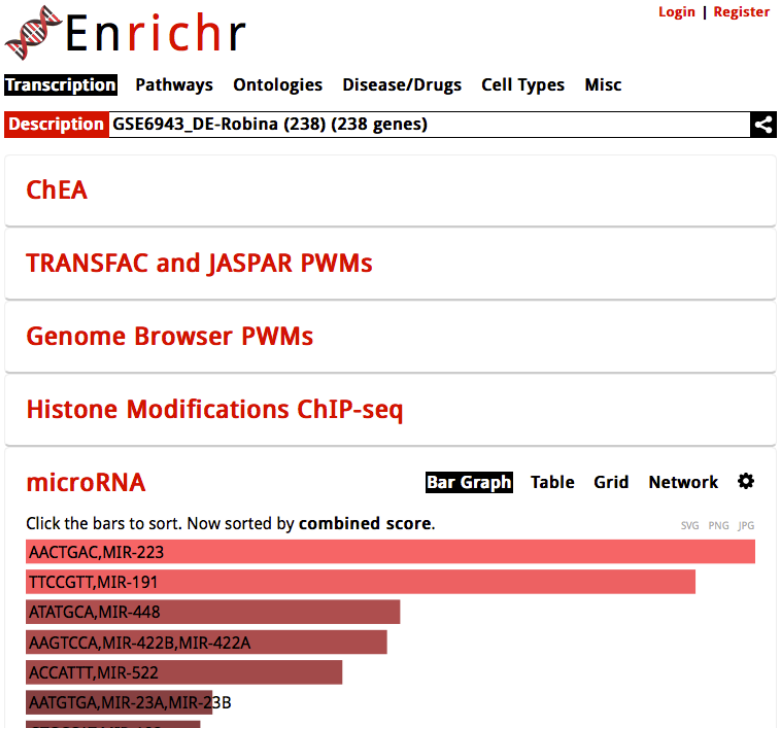
Enrichr uses a list of Entrez gene symbols as input. Each symbol in the input must be on its own line. You can upload the list by either selecting the text file that contains the list or just simply pasting in the list into the text box. We will use the gene symbols extracted from the RobiNA file, cleaned to remove duplicates, and where double-ID lines (a probeset mapping to two distinct genes) were expanded. This input file is available on the BITS server (**link (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex5-files/RobiNA-DE-genes_LFC2-FDR0.001.txt)**) and its content can be used on the **Enrich submission page (http://amp.pharm.mssm.edu/Enrichr/index.html#)**.

We present below the top part of 5 randomly selected output bar charts. please explore Enrich to get much more than only these. The online version charts are interactive and provide much more info than these pictures.

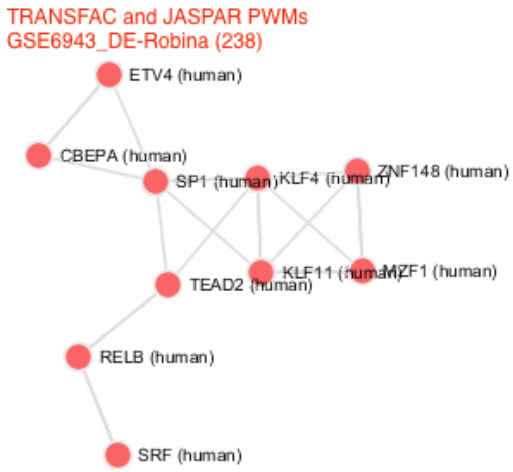



Enrichr
Login | Register
Transcription Pathways Ontologies Disease/Drugs Cell Types Misc
Description GSE6943_DE-Robina (238) (238 genes)
GO Biological Process
Bar Graph Table Grid Network
Click the bars to sort. Now sorted by combined score.
SVG PNG JPG
muscle contraction (GO:0006936)
striated muscle contraction (GO:0006941)
regulation of muscle contraction (GO:0006937)
cardiac muscle tissue morphogenesis (GO:0055008)
muscle organ development (GO:0007517)
regulation of heart contraction (GO:0008016)
calcium ion transport (GO:0006816)
regulation of striated muscle contraction (GO:0006942)
di-, tri-valent inorganic cation transport (GO:0015674)
actin filament-based movement (GO:0030048)


Enrichr
Login | Register
Transcription Pathways Ontologies Disease/Drugs Cell Types Misc
Description GSE6943_DE-Robina (238) (238 genes)
Up-regulated CMAP
Down-regulated CMAP
GeneSigDB
OMIM Disease
Bar Graph Table Grid Network
Click the bars to sort. Now sorted by combined score.
SVG PNG JPG
cardiomyopathy
myopathy
cardiomyopathy,_hypertrophic
cardiomyopathy,_dilated
muscular_dystrophy
deafness
anemia
blood
alzheimer_disease



An example of network view for the enriched Transfac & Jaspar TF-motifs



Webgestalt

This second tool largely overlaps in data-sources with **Enrich** although its tabular reporting format makes it a little less attractive to my eyes. WebGestalt accepts many ID types and allows selecting the exact background based on the array which is a plus as compared to Enrich and puts it even with DAVID in that respect.

WebGestalt [5] is a "WEB-based GEne SeT AnaLysis Toolkit". It is designed for functional genomic, proteomic and large-scale genetic studies from which large number of gene lists (e.g. differentially expressed gene sets, co-expressed gene sets etc) are continuously generated. WebGestalt incorporates information from different public resources and provides an easy way for biologists to make sense out of gene lists.

Please read the **full manual (http://bioinfo.vanderbilt.edu/webgestalt/WebGestalt_manual_2013_04_12.pdf)** for a good introduction to this tool.

The probe-list obtained from the RobiNA results is available on the BITS server (**link (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex5-files/RobiNA-DE-probes_LFC2-FDR0.001.txt)**) and its content can be used on the **WebGestalt submission page (http://bioinfo.vanderbilt.edu/webgestalt/)**.

- The ID type of the enriched list can be identified by selecting it in the list

Select the organism of interest

rnorvegicus

Select gene ID type

rnorvegicus__affy_rae230a

Upload gene list

Choose File no file selected

OR

1367707_at
1370355_at
1376371_at
1368000_at
1373410_at
1388116_at
1375230_at
1371315_at
1368966_at
1371293_at

Clear

ENTER

- In the next window, the 'Reference Set for Enrichment Analysis' should be selected from the list

Select Reference Set for Enrichment Analysis

- ✓ Select_Id_Type_from_Drop_Down_Menu
- rnorvegicus__genome
- rnorvegicus__entrezgene_protein-coding
- rnorvegicus__affy_rae230a
- rnorvegicus__affy_rae230b
- rnorvegicus__affy_raex_1_0_st_v1
- rnorvegicus__affy_ragene_1_0_st_v1
- rnorvegicus__affy_rat230_2
- rnorvegicus__affy_rg_u34a
- rnorvegicus__affy_rg_u34b
- rnorvegicus__affy_rg_u34c
- rnorvegicus__affy_rn_u34
- rnorvegicus__affy_rt_u34
- rnorvegicus__agilent_G4131A
- rnorvegicus__agilent_G4131F
- rnorvegicus__agilent_wholegenome_4x44k_v1
- rnorvegicus__agilent_wholegenome_4x44k_v3
- rnorvegicus__codelink

1388502_at
1399161_a_at
1368122_at
1368626_at
1376309_at

D Type

Significance Level

- Two sets of enrichments are available

User Data: textAreaUpload.txt. Total number of User IDs: 301. 255 user IDs can unambiguously map to 226 unique Entrez Gene IDs. 46 user IDs were mapped to multiple Entrez Gene IDs or could not be mapped to any Entrez Gene ID. The Enrichment Analysis and GO Slim Classification will be based upon the 226 unique Entrez Gene IDs.

Click here for new analysis

Enrichment Analysis

Enrichment Analysis

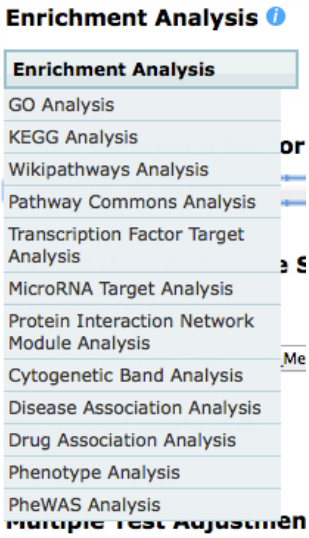
GO Slim Classification

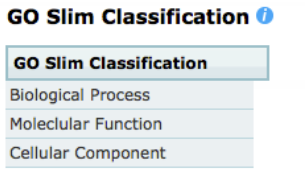
GO Slim Classification

Select Reference Set for Enrichment Analysis

rnorvegicus__affy_rae230a

-





- The results are shown in tables with annotations and scores as well as the list of genes responsible for the enrichment (KEGG pathways)

WebGestalt

WEB-based GEne SeT AnaLysis Toolkit

Translating gene lists into biological insights...

User data and parameters: User data: textAreaUpload.txt, Organism: rnorvegicus, Id Type: affy_rae230a, Ref Set: affy_rae230a, Significance Level: Top10, Statistics Test: Hypergeometric, MTC: BH, Minimum: 2

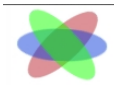
This table lists the enriched KEGG pathways, number of Entrez IDs in your user data set for the pathway, the corresponding Entrez IDs, and the statistics for the enriched pathway.
The statistic column lists:

- C: the number of reference genes in the category
- O: the number of genes in the gene set and also in the category
- E: the expected number in the category
- R: ratio of enrichment
- rawP: p value from hypergeometric test
- adjP: p value adjusted by the multiple test adjustment

Finally, the pathway name is linked to KEGG where the user ids are highlighted, the number of user gene ids is linked to a table with information about the user ids, and the Entrez IDs are linked to Entrez Gene.

PathwayName	#Gene	EntrezGene	Statistics
Dilated cardiomyopathy	15	682930 24837 29658 29275 81636 29248 295929 29556 25399 117557 24239 64532 689560 64672 116600	C=70;O=15;E=1.54;R=9.72;rawP=2.19e-11;adjP=1.58e-09
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	12	682930 29658 171009 83501 306871 287925 25399 24239 689560 24392 307505 116600	C=58;O=12;E=1.28;R=9.38;rawP=3.59e-09;adjP=1.29e-07
Hypertrophic cardiomyopathy (HCM)	12	682930 24837 29658 29275 29248 295929 29556 25399 117557 24239 689560 116600	C=66;O=12;E=1.46;R=8.25;rawP=1.70e-08;adjP=4.08e-07
Cardiac muscle contraction	12	682930 24837 29658 29275 29248 84396 29556 25399 117557 24239 689560 116600	C=70;O=12;E=1.54;R=7.78;rawP=3.39e-08;adjP=6.10e-07
Tight junction	11	83807 171009 24584 289759 691644 85420 29556 360543 81755 287408 307505	C=101;O=11;E=2.23;R=4.94;rawP=1.30e-05;adjP=0.0002
Vascular smooth muscle contraction	10	682930 81636 85420 362039 58965 24239 29354 64532 24173 117558	C=93;O=10;E=2.05;R=4.88;rawP=3.63e-05;adjP=0.0004
Calcium signaling pathway	13	682930 81636 116601 296369 64561 24239 689560 29353 24245 24173 117558 64672 114207	C=162;O=13;E=3.57;R=3.64;rawP=5.87e-05;adjP=0.0006
GnRH signaling pathway	9	682930 81636 60352 362039 24239 29354 64532 114495 24245	C=82;O=9;E=1.81;R=4.98;rawP=7.60e-05;adjP=0.0007
Pancreatic secretion	9	54242 81779 81636 29354 689560 64532 116601 84396 362039	C=85;O=9;E=1.87;R=4.80;rawP=0.0001;adjP=0.0008
Insulin signaling pathway	8	25058 64561 50671 114508 689995 114203 29353 25739	C=113;O=8;E=2.49;R=3.21;rawP=0.0035;adjP=0.0252

- (Wiki pathways)



WEB-based GEne SeT AnaLysis Toolkit

Translating gene lists into biological insights...

User data and parameters: User data: textAreaUpload.txt, Organism: rnorvegicus, Id Type: affy_rae230a, Ref Set: affy_rae230a, Significance Level: Top10, Statistics Test: Hypergeometric, MTC: BH, Minimum: 2

This table lists the enriched Wikipathways, number of Entrez IDs in your user data set for the pathway, the corresponding Entrez IDs, and the statistics for the enriched pathway.
The statistic column lists:

- C: the number of reference genes in the category
- O: the number of genes in the gene set and also in the category
- E: the expected number in the category
- R: ratio of enrichment
- rawP: p value from hypergeometric test
- adjP: p value adjusted by the multiple test adjustment

Finally, gene set is linked to the wikipathway graph, which is generated dynamically from bioinfo.vanderbilt.edu/wg_gsat, the number of user gene ids is linked to a table with information about the user ids, and the Entrez IDs are linked to Entrez Gene.

PathwayName	#Gene	EntrezGene	Statistics
Striated Muscle Contraction	16	311029 171009 29388 24838 29275 295929 29556 296369 362867 117557 56781 24837 29389 29248 292879 171409	C=35;O=16;E=0.77;R=20.74;rawP=5.13e-18;adjP=1.95e-16
Calcium Regulation in the Cardiac Cell	12	85420 60449 64532 24392 64672 686019 682930 81636 24239 689560 24245 24173	C=132;O=12;E=2.91;R=4.12;rawP=3.34e-05;adjP=0.0006
Myometrial Relaxation and Contraction Pathways	10	29275 85420 60449 64532 24392 117558 81636 58965 689560 24245	C=134;O=10;E=2.95;R=3.38;rawP=0.0007;adjP=0.0089
Glycogen Metabolism	4	64561 29353 24645 25739	C=32;O=4;E=0.71;R=5.67;rawP=0.0051;adjP=0.0485
Diurnally regulated genes with circadian orthologs	4	24253 305234 498642 25714	C=43;O=4;E=0.95;R=4.22;rawP=0.0145;adjP=0.1102
Glucuronidation	2	24645 25058	C=10;O=2;E=0.22;R=9.07;rawP=0.0194;adjP=0.1229
Integrin-mediated cell adhesion	5	60352 301509 29155 114495 117558	C=80;O=5;E=1.76;R=2.83;rawP=0.0315;adjP=0.1496
G13 Signaling Pathway	3	366624 56781 83708	C=30;O=3;E=0.66;R=4.54;rawP=0.0277;adjP=0.1496
Glycolysis and Gluconeogenesis	3	25438 114508 25058	C=38;O=3;E=0.84;R=3.58;rawP=0.0507;adjP=0.1991
SIDS Susceptibility Pathways	4	29253 25665 60449 689560	C=64;O=4;E=1.41;R=2.83;rawP=0.0524;adjP=0.1991

Again, many more tables can be generated in WebGestalt and you should choose the type of enrichment that fits your experimental needs. Data can be saved back to disk for further use.

Conclusion

More complete analysis can be performed by those few who can program in [R]-Bioconductor. As this session is aimed at Biologists, this option is not further discussed. Users can also consider using commercial packages like Ingenuity pathway Analysis that provide much more detailed and rich information than what free tools can offer.

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑ http://david.abcc.ncifcrf.gov
2. ↑ http://www.nature.com/nprot/journal/v4/n1/abs/nprot.2008.211.html

   Da Wei Huang, Brad T Sherman, Richard A Lempicki
   Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.
   Nat Protoc: 2009, 4(1);44-57
   [PubMed:19131956] ##WORLDCAT## [DOI] (I p)

3. ↑ http://amp.pharm.mssm.edu/Enrichr/

   Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, Avi Ma'ayan
   Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.
   BMC Bioinformatics: 2013, 14;128
   [PubMed:23586463] ##WORLDCAT## [DOI] (I e)

4. ↑

   Alexander Lachmann, Avi Ma'ayan
   Lists2Networks: integrated analysis of gene/protein lists.
   BMC Bioinformatics: 2010, 11;87
   [PubMed:20152038] ##WORLDCAT## [DOI] (I e)

5. ↑ http://bioinfo.vanderbilt.edu/webgestalt/

   Stefan Kirov, Ruiru Ji, Jing Wang, Bing Zhang
   Functional annotation of differentially regulated gene set using WebGestalt: a gene set predictive of response to ipilimumab in tumor biopsies.
   Methods Mol. Biol.: 2014, 1101;31-42
   [PubMed:24233776] ##WORLDCAT## [DOI] (I p)

   Jing Wang, Dexter Duncan, Zhiao Shi, Bing Zhang
   WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013.
   Nucleic Acids Res.: 2013, 41(Web Server issue);W77-83
   [PubMed:23703215] ##WORLDCAT## [DOI] (I p)

   Bing Zhang, Stefan Kirov, Jay Snoddy
   WebGestalt: an integrated system for exploring gene sets in various biological contexts.
   Nucleic Acids Res.: 2005, 33(Web Server issue);W741-8
   [PubMed:15980575] ##WORLDCAT## [DOI] (I p)

```
http://bioinfo.vanderbilt.edu/webgestalt/WebGestalt_manual_2013_04_12.pdf
```

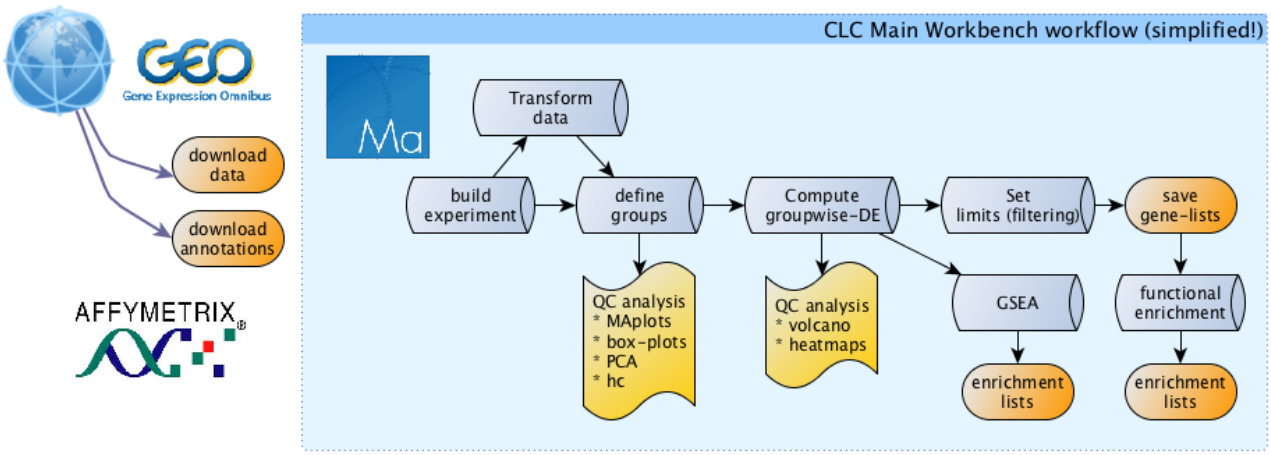
Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=PubMA_Exercise.5&oldid=10960"
Category: PUBMA2014

- This page was last modified on 26 August 2014, at 14:09.
- This page has been accessed 150 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

PubMA Exercise.6

From BioWareWIKI

Full analysis workflow using CLC main workbench



The following content was directly taken from the current CLC documentation (Feb17, 2014) and applies only for people with access to the CLC Main workbench

Contents

- 1 CLC Tutorial material
  - 1.1 Loading microarray data into the Workbench
  - 1.2 Loading your own Affymetrix microarray data into the Workbench
  - 1.3 Main results and specific settings
  - 1.4 Enrichment analysis in CLC Main
    - 1.4.1 hypergeometric tTest
    - 1.4.2 GSEA
- 2 Published results
- 3 download exercise files

CLC Tutorial material

Users from VIB labs or people having a license for the CLC Main workbench can proceed with this exercise during the afternoon open session or later from home. The final CLC project folder can be downloaded from the BITS server ('Heart vs Diaphragm.zip') and imported in the User CLC project-manager. The workflow is not further discussed here and we invite interested people to follow the four PDF files also present on the server.

- Expression_analysis_part_I.pdf (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex6-files/Expression_analysis_part_I.pdf)
- Expression_analysis_part_II.pdf (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex6-files/Expression_analysis_part_II.pdf)
- Expression_analysis_part_III.pdf (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex6-files/Expression_analysis_part_III.pdf)
- Expression_analysis_part_IV.pdf (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex6-files/Expression_analysis_part_IV.pdf)

A recompiled version of this tutorial is part of a former BITS CLC-training accessible **Here (http://data.bits.vib.be/pub/trainingen/CLCMain/TutorialMicroarrays.pdf)**.

Loading microarray data into the Workbench

The Workbench supports analysis of **one-color** expression arrays. These may be imported from GEO (http://www.ncbi.nlm.nih.gov/). The Workbench supports the following formats:

- GEO SOFT sample-files (http://www.ncbi.nlm.nih.gov/geo/info/soft2.html): simple line-based, plain text files that contain all the data and the descriptive information of a microarray experiment (example SOFT-file (http://www.ncbi.nlm.nih.gov/geo/info/soft_ex_platform.txt))
- GEO series-file: txt-files containing the definitions of a group of related samples. They contain tables describing extracted data, summary conclusions and analyses. Each Series-file is assigned a unique GEO accession number (GSExxx).

Loading your own Affymetrix microarray data into the Workbench

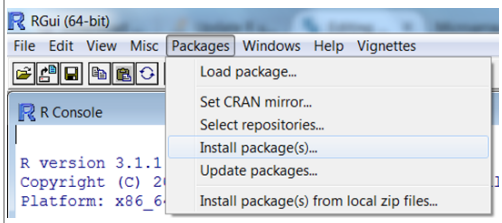
The Workbench assumes that expression values are given at the gene (probe set) level, thus probe-level analysis of Affymetrix arrays and import of Affymetrix CEL- and CDF-files is not supported. However, you can import your own Affymetrix data via two ways:

- as .CHP files generated by Affymetrix Expression Console containing normalized Affymetrix data
  See the section on how to convert .CEL files to .CHP files using the Expression Console (http://wiki.bits.vib.be/index.php/Analyze_GEO_data_with_the_Affymetrix_software#Converting_CEL_data_to_CHP_format_required_for_TAC) for a detailed discussion on how to do this. Use RMA for the normalisation !
- as .txt files exported from R containing normalized Affymetrix data

Expand this section to see how you can do the normalization in R. [hide]

To create these txt files, open R (http://www.r-project.org) or RStudio (http://www.rstudio.com/) as administrator and install the following packages:

- Matrix
- lattice
- fdrtool
- rpart



Then run the following code:

```
# Install all required Bioconductor packages
source("http://bioconductor.org/biocLite.R")
biocLite()
biocLite("affy")
```

You also need Bioconductor packages for the annotation of the spots on the arrays. In the example below I use the annotation packages for the Arabidopsis ATH1 array. Of course, you need the annotation packages that correspond to the arrays that you used in your experiment.
There are two possibilities for obtaining these packages:

- Bioconductor has a list of annotation packages (http://www.bioconductor.org/packages/release/data/annotation/) (generated by Affymetrix). In this list find the name of the cdf file that corresponds to the array that you have used (we used ATH1 arrays so the cdf file is called ath1121501cdf):

Package	Maintainer	Title
adme16cod.db	Diego Diez	Codelink ADME Rat 16-Assay Bioarray annotation data (chip adme16cod)
ag.db	Bioconductor Package Maintainer	Affymetrix Arabidopsis Genome Array annotation data (chip ag)
agcdf	Bioconductor Package Maintainer	agcdf
agprobe	Bioconductor Package Maintainer	Probe sequence data for microarrays of type ag
anopheles.db0	Bioconductor Package Maintainer	Base Level Annotation databases for anopheles
arabidopsis.db0	Bioconductor Package Maintainer	Base Level Annotation databases for arabidopsis
ath1121501.db	Bioconductor Package Maintainer	Affymetrix Arabidopsis ATH1 Genome Array annotation data (chip ath1121501)
ath1121501cdf	Bioconductor Package Maintainer	ath1121501cdf
ath1121501probe	Bioconductor Package Maintainer	Probe sequence data for microarrays of type ath1121501

  To use this cdf file execute the following R-commands:

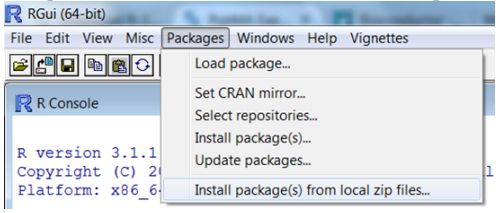
```
biocLite("ath1121501cdf")

# load Bioconductor libraries
library("ath1121501cdf")
library("affy")

# specify path on your computer where the directory that contains the CEL-files is located
celpath = "C:/Users/Janick/My Documents/R/win-library/2.14/affydata/celfiles/"

# import CEL files containing raw probe-level data
data = ReadAffy(celfile.path=celpath)
```

- BrainArray provides a list of custom annotation packages (http://brainarray.mbni.med.umich.edu/bioc/bin/windows/contrib/3.0/). To use these cdf files, download the zip file from the website. Install it from the local zip file:



  Then execute the following code:

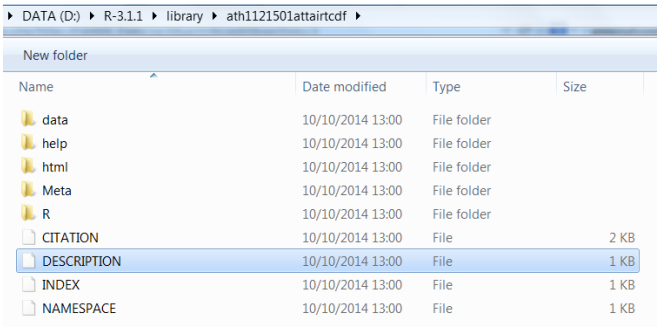
```
# load Bioconductor libraries
library("affy")
```

```
# specify path on your computer where the directory that contains the CEL-files is located
celpath = "D:/R-2.15.2/library/affydata/celfiles/Apum/"

# import CEL files containing raw probe-level data
data = ReadAffy(celfile.path=celpath)

# indicate you want to use the custom cdf
# If you don't specify the cdfname, Bioconductor will use the default Affymetrix cdf.
data@cdfName="ATH1121501_At_TAIRT"
```

You can find the name of the custom cdf by going to the package folder:


DATA (D:) ▸ R-3.1.1 ▸ library ▸ ath1121501attairtcdf ▸
New folder
Name
Date modified
Type
Size
data 10/10/2014 13:00 File folder
help 10/10/2014 13:00 File folder
html 10/10/2014 13:00 File folder
Meta 10/10/2014 13:00 File folder
R 10/10/2014 13:00 File folder
CITATION 10/10/2014 13:00 File 2 KB
DESCRIPTION 10/10/2014 13:00 File 1 KB
INDEX 10/10/2014 13:00 File 1 KB
NAMESPACE 10/10/2014 13:00 File 1 KB

Open the **DESCRIPTION** file and look in the **Description** line:

```
Package: ath1121501attairtcdf
Title: ath1121501attairtcdf
Version: 18.0.0
Created: Wed Jan 29 12:33:39 2014
Author: Manhong Dai
Description: A package containing an environment representing the
customcdf_out/tairt.out/ATH1121501_At_TAIRT.cdf file.
Maintainer: Manhong Dai <daimh@umich.edu>
License: LGPL
biocViews: MBNICustomCDF, AnnotationData, AffymetrixChip, ath1121501,
        ath1121501attairt, Arabidopsis_thaliana
```

Then proceed by executing the following R-code:

```
# normalisation using 'rma' algorithm.
data.rma=rma(data)
# The output is an exprSet object with a data matrix containing normalized log-intensities (on probe set-level) in the exprs slot.

# writing probe-set level data to a file called 'data.txt'
write.exprs(data.rma,file="data.txt")
```

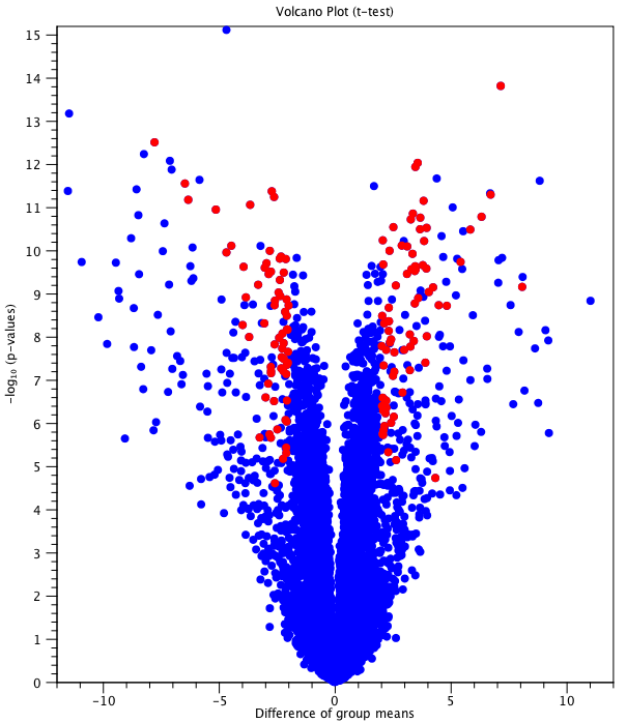
The resulting text file can be imported into the Workbench.

Main results and specific settings

Only key steps are reproduced here to provide information to interpret the figures. All other steps and parameters will be explained in the tutorial PDF files linked above.

After computing group-wise differential expression, a filtering step is applied to the full table to retain only DE genes with at least 2-fold change in expression, with an adjusted p-value of at most 5x10-3 and with expression data (present calls) in at least 4 of the 6 replicates.

The classical volcano plot is produced with in red the 142 DE genes selected during filtering. This subset will be used as 'test' set against all other genes in the data table in the hypergeometric enrichment analyses detailed below.

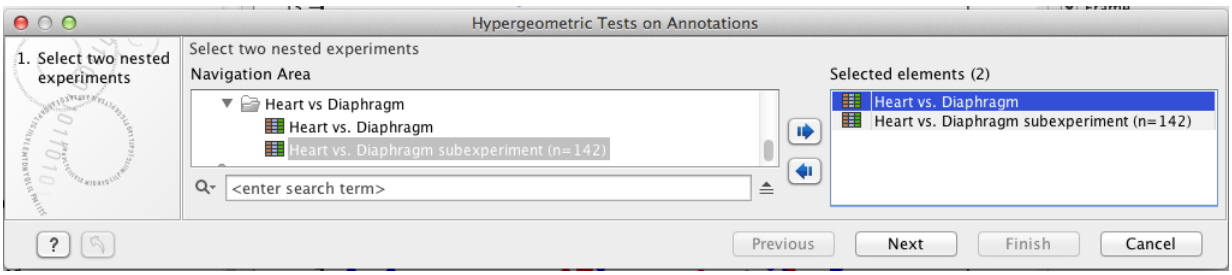


Due to the logarithmic nature of the data (transformed), the 'Difference' column should be used instead of 'Fold-Change' to represent the differential expression

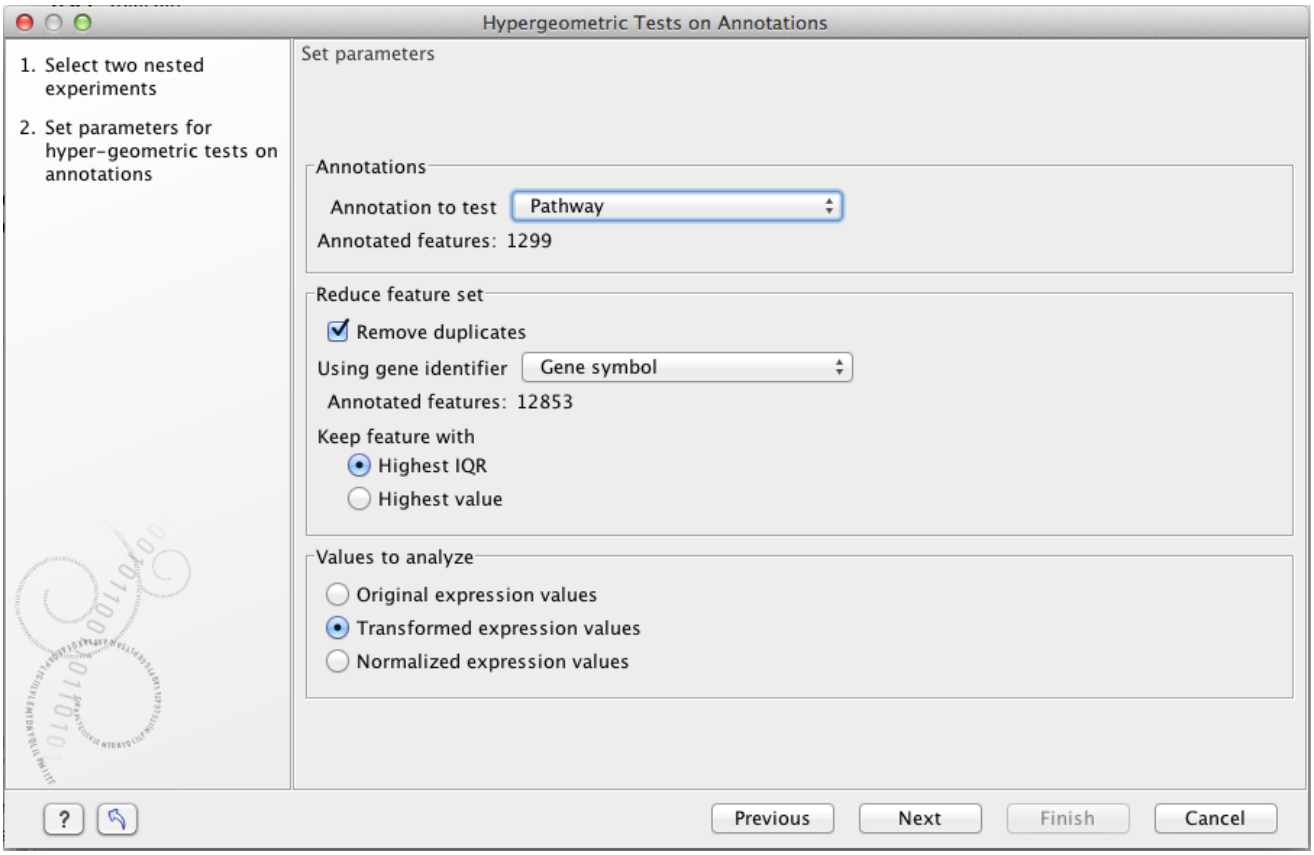
Enrichment analysis in CLC Main

hypergeometric tTest

The following figure shows data-samples used in the **hypergeometric tTest**



A window allows selecting the annotation type to be used in the test and the action to take with duplicate probes (here: 'merged by gene symbol to the highest IQR').



Hyper-geometric test for association of annotation categories to a sublist of a larger gene list (Wed Aug 13 12:38:46 CEST 2014)
Version: CLC Main Workbench 7.0.3
User: splaisan
Parameters:

```
Gene identifier column used in tests  = Gene symbol
Annotation column used in tests  = GO biological process
Raw universal gene list size  = 15923
Used universal gene list size (requiring annotation and one feature per gene only)  = 8603
Raw subset gene list size  = 142
Used subset gene list size (requiring annotation)  = 73
Expression values used when filtering to one feature per gene  = Transformed expression values
Applied filter to reduce features to one per gene  = true
Filter applied to reduce features to one per gene  = used feature with highest IQR
```

The next figure details the **Top-30** hypergeometric results for the contrast **Heart-vs-diaphragm**

	A	B	C	D	E	F	G
1	Categor	Description	Full se	In subs	Expected in subs	Observed expect	p-valu
2	5977	glycogen metabolic process	27	7	0	7	1,84E-09
3	5975	carbohydrate metabolic process	129	7	1	6	0,000104
4	6874	cellular calcium ion homeostasis	64	5	1	4	0,000199
5	60048	cardiac muscle contraction	36	4	0	4	0,00023
6	7519	skeletal muscle tissue development	42	4	0	4	0,00042
7	10890	positive regulation of sequestering of triglyceride	5	2	0	2	0,0007
8	32780	negative regulation of ATPase activity	7	2	0	2	0,001454
9	35914	skeletal muscle cell differentiation	32	3	0	3	0,002443
10	55003	cardiac myofibril assembly	10	2	0	2	0,003064
11	55008	cardiac muscle tissue morphogenesis	10	2	0	2	0,003064
12	43268	positive regulation of potassium ion transport	12	2	0	2	0,004445
13	60047	heart contraction	13	2	0	2	0,005224
14	10880	regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum	13	2	0	2	0,005224
15	6807	nitrogen compound metabolic process	13	2	0	2	0,005224
16	48738	cardiac muscle tissue development	13	2	0	2	0,005224
17	5978	glycogen biosynthetic process	14	2	0	2	0,006061
18	60025	regulation of synaptic activity	1	1	0	1	0,008493
19	6542	glutamine biosynthetic process	1	1	0	1	0,008493
20	1901020	negative regulation of calcium ion transmembrane transporter activity	1	1	0	1	0,008493
21	86023	adrenergic receptor signaling pathway involved in heart process	1	1	0	1	0,008493
22	86092	regulation of the force of heart contraction by cardiac conduction	1	1	0	1	0,008493
23	6603	phosphocreatine metabolic process	1	1	0	1	0,008493
24	42823	pyridoxal phosphate biosynthetic process	1	1	0	1	0,008493
25	2634	regulation of germinal center formation	1	1	0	1	0,008493
26	35425	autocrine signaling	1	1	0	1	0,008493
27	31275	regulation of lateral pseudopodium assembly	1	1	0	1	0,008493
28	51684	maintenance of Golgi location	1	1	0	1	0,008493
29	46439	L-cysteine metabolic process	1	1	0	1	0,008493
30	42424	catecholamine catabolic process	1	1	0	1	0,008493

Another page presents details about the parameters used in the different tests

- results of the hypergeometric test for GO-BP

Hyper-geometric test for association of annotation categories to a sublist of a larger gene list (Wed Aug 13 14:49:36 CEST 2014)
Version: CLC Main Workbench 7.0.3
User: splaisan
Parameters:

```
Gene identifier column used in tests  = Gene symbol
Annotation column used in tests  = GO molecular function
Raw universal gene list size  = 15923
Used universal gene list size (requiring annotation and one feature per gene only)  = 8969
Raw subset gene list size  = 142
Used subset gene list size (requiring annotation)  = 74
Expression values used when filtering to one feature per gene  = Transformed expression values
Applied filter to reduce features to one per gene  = true
Filter applied to reduce features to one per gene  = used feature with highest IQR
```

- results of the hypergeometric test for GO-MF

Hyper-geometric test for association of annotation categories to a sublist of a larger gene list (Wed Aug 13 14:49:36 CEST 2014)
Version: CLC Main Workbench 7.0.3
User: splaisan
Parameters:

```
Gene identifier column used in tests  = Gene symbol
Annotation column used in tests  = GO molecular function
Raw universal gene list size  = 15923
Used universal gene list size (requiring annotation and one feature per gene only)  = 8969
Raw subset gene list size  = 142
Used subset gene list size (requiring annotation)  = 74
Expression values used when filtering to one feature per gene  = Transformed expression values
Applied filter to reduce features to one per gene  = true
Filter applied to reduce features to one per gene  = used feature with highest IQR
```

- results of the hypergeometric test for Pathways

Hyper-geometric test for association of annotation categories to a sublist of a larger gene list (Wed Aug 13 14:49:36 CEST 2014)
Version: CLC Main Workbench 7.0.3
User: splaisan
Parameters:

```
Gene identifier column used in tests  = Gene symbol
Annotation column used in tests  = GO molecular function
Raw universal gene list size  = 15923
Used universal gene list size (requiring annotation and one feature per gene only)  = 8969
Raw subset gene list size  = 142
Used subset gene list size (requiring annotation)  = 74
Expression values used when filtering to one feature per gene  = Transformed expression values
Applied filter to reduce features to one per gene  = true
Filter applied to reduce features to one per gene  = used feature with highest IQR
```

GSEA

The **GSEA method** does not require partitioning the data as for the hypergeometric test, it takes the full table and considers the relative ranking of gene list members in relation to the individual gene expression levels in the data.

- Settings for the GSEA test for GO-BP

Gene set enrichment analysis (Wed Aug 13 12:42:16 CEST 2014)
Version: CLC Main Workbench 7.0.3
User: splaisan
Parameters:

```
Gene identifier column used in tests  = Gene symbol
Annotation column used in tests  = GO biological process
The features were ranked on  = t-statistic
group comparison  = Heart - Diaphragm
Raw universal gene list size  = 15923
Used universal gene list size (requiring annotation and one feature per gene only)  = 8603
Applied filter to reduce features to one per gene  = true
Filtered features to one-per-gene by using  = feature with highest IQR
Expression values used when filtering features  = Transformed expression values
Minimum category size required  = 10
Number categories passing mimimum  size requirement = 1451
Number permutations for p-value calculation:  = 10000
```

- Settings for the GSEA test for GO-MF

Gene set enrichment analysis (Wed Aug 13 14:51:03 CEST 2014)
Version: CLC Main Workbench 7.0.3
User: splaisan
Parameters:

```
Gene identifier column used in tests  = Gene symbol
Annotation column used in tests  = GO molecular function
The features were ranked on  = t-statistic
group comparison  = Heart - Diaphragm
Raw universal gene list size  = 15923
Used universal gene list size (requiring annotation and one feature per gene only)  = 8969
Applied filter to reduce features to one per gene  = true
Filtered features to one-per-gene by using  = feature with highest IQR
Expression values used when filtering features  = Transformed expression values
Minimum category size required  = 10
Number categories passing mimimum  size requirement = 503
Number permutations for p-value calculation:  = 10000
```

- Settings for the GSEA test for Pathways

Gene set enrichment analysis (Wed Aug 13 15:10:25 CEST 2014)
Version: CLC Main Workbench 7.0.3
User: splaisan
Parameters:

```
Gene identifier column used in tests  = Gene symbol
Annotation column used in tests  = Pathway
The features were ranked on  = t-statistic
group comparison  = Heart - Diaphragm
Raw universal gene list size  = 15923
Used universal gene list size (requiring annotation and one feature per gene only)  = 994
Applied filter to reduce features to one per gene  = true
Filtered features to one-per-gene by using  = feature with highest IQR
Expression values used when filtering features  = Transformed expression values
Minimum category size required  = 10
Number categories passing mimimum  size requirement = 45
Number permutations for p-value calculation:  = 10000
```

Top-10 GSEA results for BP increased in Heart

	A	B	C	D	E	F
1	Catego	Description	Size	Test statist	Lower tai	Upper tai
2	7519	skeletal muscle tissue development	42	-28,138605	0	1
3	3009	skeletal muscle contraction	15	-32,143862	0	1
4	6936	muscle contraction	37	-43,126281	0	1
5	5977	glycogen metabolic process	27	-18,845885	0,0002	0,9998
6	6094	gluconeogenesis	23	-18,815161	0,0002	0,9998
7	6937	regulation of muscle contraction	19	-21,287718	0,0002	0,9998
8	5978	glycogen biosynthetic process	14	-21,712337	0,0003	0,9997
9	6412	translation	205	-14,686305	0,0004	0,9996
10	6413	translational initiation	46	-18,199774	0,0004	0,9996

Top-10 GSEA results for BP increased in Diaphragm

	A	B	C	D	E	F
1	Catego	Description	Size	Test statist	Lower tai	Upper tai
2	55114	oxidation-reduction process	473	28,5285603	1	0
3	86005	regulation of ventricular cardiac muscle cell action potential	13	28,4480485	1	0
4	8016	regulation of heart contraction	25	21,4227626	0,9999	0,0001
5	55010	ventricular cardiac muscle tissue morphogenesis	25	18,9272308	0,9996	0,0004
6	6099	tricarboxylic acid cycle	24	17,8151236	0,9994	0,0006
7	86004	regulation of cardiac muscle cell contraction	12	18,7923644	0,9993	0,0007
8	86091	regulation of heart rate by cardiac conduction	15	18,0252284	0,9993	0,0007
9	51291	protein heterooligomerization	82	16,8587409	0,9992	0,0008
10	2026	regulation of the force of heart contraction	21	17,9066909	0,9992	0,0008

Published results

The original publication ends with functional enrichment results identifying key differences between 'heart' and 'diaphragm' tissues in rats. We link here to the results published by 'van Lunteren E, Spiegler S, Moyer M' [1]. Full details about this dataset can be found on the **http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=+GSE6943** GEO page.

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑

   Erik van Lunteren, Sarah Spiegler, Michelle Moyer
   Contrast between cardiac left ventricle and diaphragm muscle in expression of genes involved in carbohydrate and lipid metabolism.
   Respir Physiol Neurobiol: 2008, 161(1);41-53
   [PubMed:18207466] ##WORLDCAT## [DOI] (P p)

   http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=+GSE6943

[Main_Page | Hands-on Analysis of public microarray datasets | PubMA_Exercise.5 | **PubMA_Exercise.6** |
| Analyze GEO data with the Affymetrix software]

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=PubMA_Exercise.6&oldid=11808"
Category: PUBMA2014

- This page was last modified on 20 October 2014, at 16:21.
- This page has been accessed 88 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

Analyze GEO data with the Affymetrix software

From BioWareWIKI

Analyzing a selected GEO dataset using the Affymetrix Expression Console (EC) and Transcriptome Analysis Console (TAC)

Contents

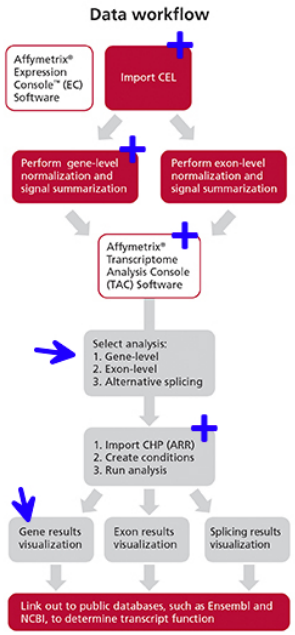
- 1 Introduction
- 2 The Affymetrix Expression Console (EC)
  - 2.1 Converting CEL data to CHP format required for TAC
  - 2.2 Performing QC on the data and generating summarizing plots
- 3 The Affymetrix Transcriptome Analysis Console (TAC)
  - 3.1 Importing EC data and defining Groups
  - 3.2 Computing 'gene-level' Differential expression
  - 3.3 Adjusting Differential expression limits
  - 3.4 Plots based on the filtered differential expression table
  - 3.5 Exporting results
- 4 Conclusion
- 5 Youtube videos from the Affymetrix training team
- 6 download exercise files

Introduction

*The data used in this how-to tutorial is the same as that used for the BITS hands-on training **Hands-on_Analysis_of_public_microarray_datasets***

The Affymetrix online training page dedicated to MA and transcriptome analysis can be browsed **here (http://www.affymetrix.com/estore/browse/level_seven_software_products_only.jsp?productId=131414&categoryId=35623&productName=Affymetrix%2526%2523174%253B-Expression-Console%2526%2523153%253B-Software#1_1)**[1]; This main pages contains links to download the necessary software as well as links to other Affymetricx resources necessary to perform a full expression analysis. Also refer to the Affymetrix **Transcriptome Analysis Console (TAC) Software and Expression Console Software tutorial pages (http://www.affymetrix.com/support/learning/training_tutorials/tac_ec/index.affx#1_2)** [2]

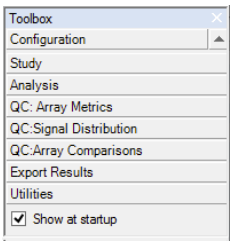
*You will need to set a free **NetAffx (http://www.affymetrix.com/analysis/index.affx)** account to download software and access data pages*


Data workflow
Affymetrix® Expression Console™ (EC) Software
Import CEL
Perform gene-level normalization and signal summarization
Perform exon-level normalization and signal summarization
Affymetrix® Transcriptome Analysis Console (TAC) Software
Select analysis:
1. Gene-level
2. Exon-level
3. Alternative splicing
1. Import CHP (ARR)
2. Create conditions
3. Run analysis
Gene results visualization
Exon results visualization
Splicing results visualization
Link out to public databases, such as Ensembl and NCBI, to determine transcript function

A summarized in the above picture, we will now perform the two steps required to perform a full analysis starting from a set of CEL files obtained from the GEO repository. The method can be divided into two steps as detailed below; the first step converts CEL data to a format better suited for differential expression analysis using the Expression console; the second step computes differential expression base don user-defined sample groups and using the Transcription Analysis Console. Results presented here correspond to the blue-highlights in the above workflow.

The Affymetrix Expression Console (EC)

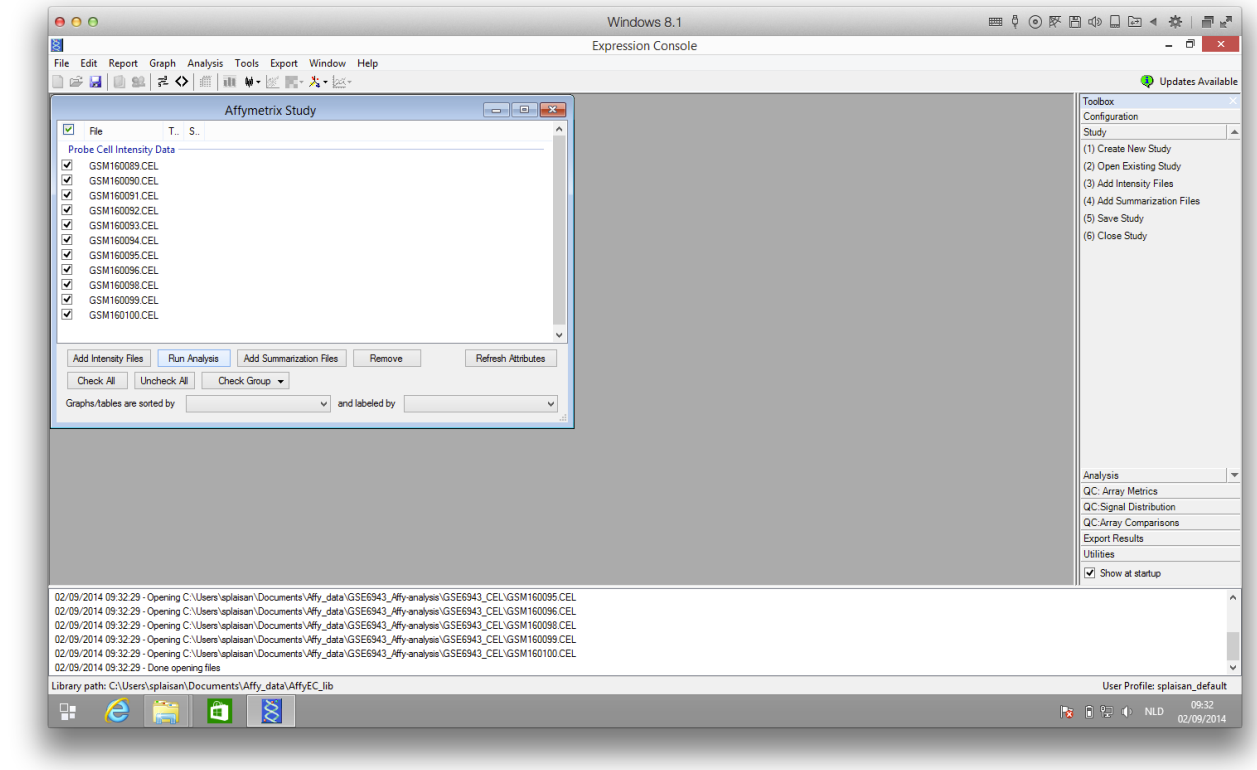
The EC software allows step by step processing of the data by sequentially clicking each tool on the right hand side of the window



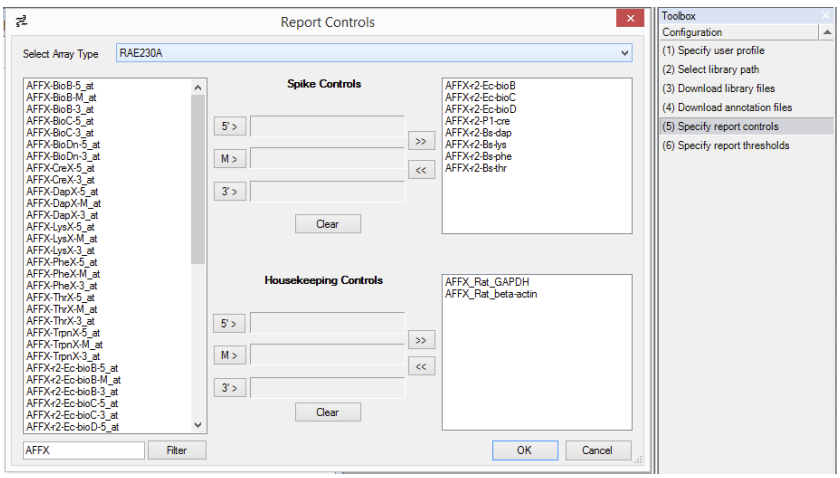
Other 'Configuration' tools are not detailed here.

Converting CEL data to CHP format required for TAC

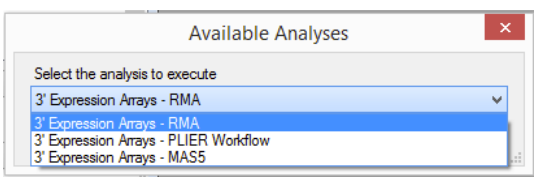
Using the 'Study' tools, the CEL files downloaded from GEO are loaded in the software, then normalized using a chosen method (out of RMA, MAS5 and PLIER). We use RMA as this is the standard method.



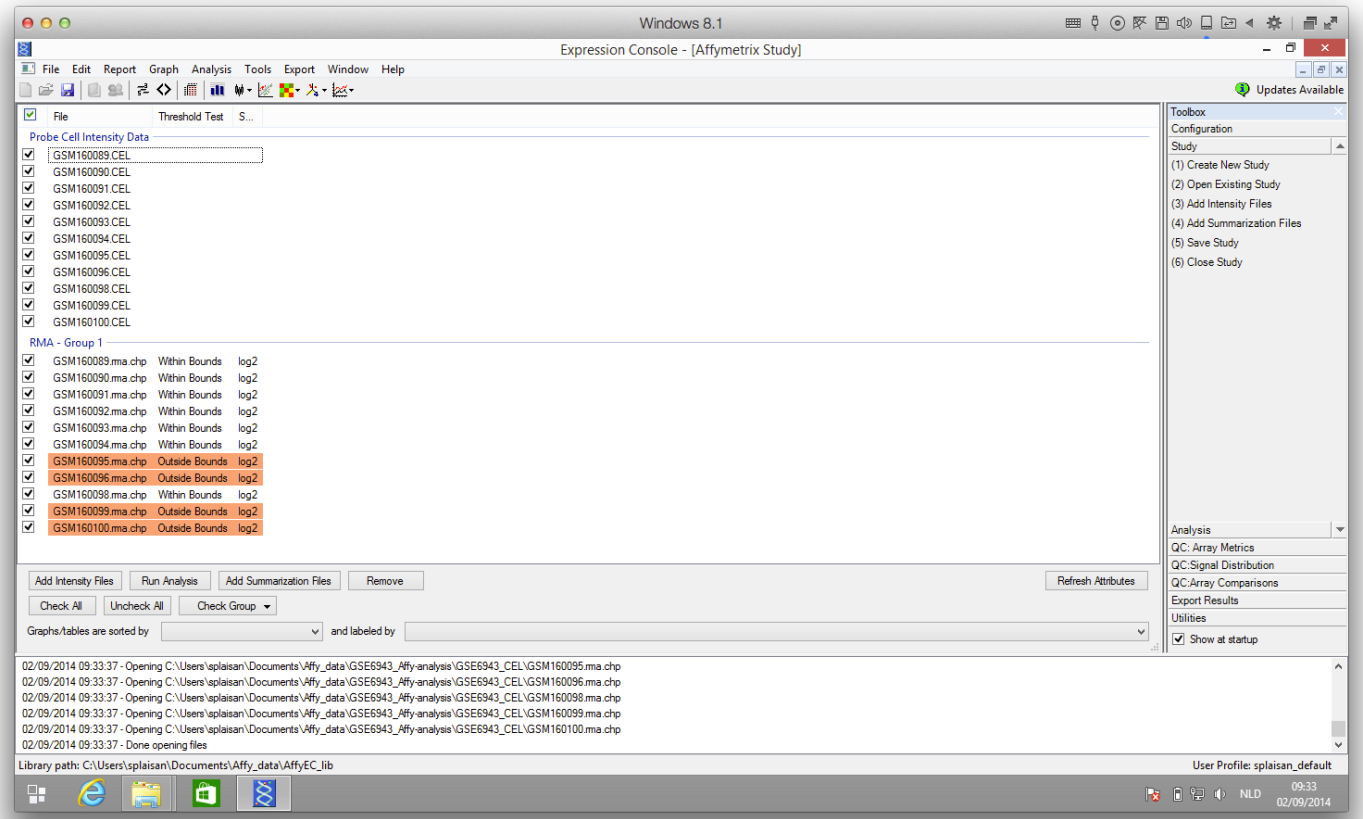
The interface allowing defining data quality controls used by the software can be reached from the right workflow items 'report-controls'.



The choice of the right method to apply for normalization is not detailed here, please refer to the BITS microarray training session and material for more information about this topic (**Introduction to Affymetrix Microarray Anaysis (https://www.bits.vib.be/index.php/bits-search-results?searchword=Intro%20Affymetrix&searchphrase=all)**). The normalization method is selected from a pop-down menu.


Available Analyses
Select the analysis to execute
3' Expression Arrays - RMA
3' Expression Arrays - RMA
3' Expression Arrays - PLIER Workflow
3' Expression Arrays - MAS5

The process takes some time and leads to a summary page and saves new files to the disk with extension '.chp' containing the normalized data, one for each imported CEL file. The '.chp' files are ready for import in the TAC tool



As seen above, several samples are reported 'outside bounds' by the RMA workflow. It means that some control probe sets did not meet the quality requirements. We looked it up and saw that the sample prep control probe sets (targeting B. subtilis genes: dap, thr, phe and lys) were not behaving as expected. Dap RNA is added in higher concentrations than thr RNA so the signal of dap should be higher than that of thr and this was not the case for the samples that were flagged 'outside bounds'. The other control probes behaved as they should. So it might be that in some samples the reverse transcription of the high abundance transcripts was not completely efficient (because of saturation...).

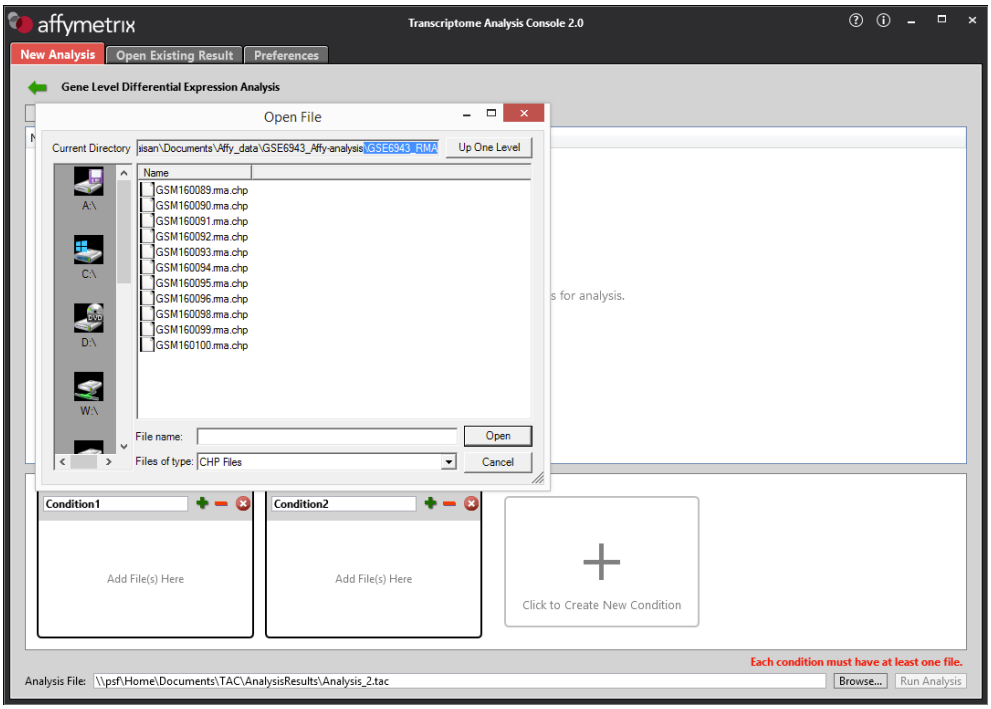
As part of the standard Affymetrix microarray processing, control molecules are added to the mRNA at different concentrations prior to producing the cDNA. Other molecules (cDNA) are added later in the sample preparation to control for hybridization on the chip. The out of bound errors reported above result from the discrepancy between the known spiked-in quantities and the readout after scanning the chip. The highest concentration of control does not produce a final value higher than a lower concentration of control which results in raising an alarm and showing the 4 samples with colored background. Full details about the identity of the faulty probes and the obtained values can be found at the bottom table part of the full report linked in the next paragraph (PDF)

Performing QC on the data and generating summarizing plots

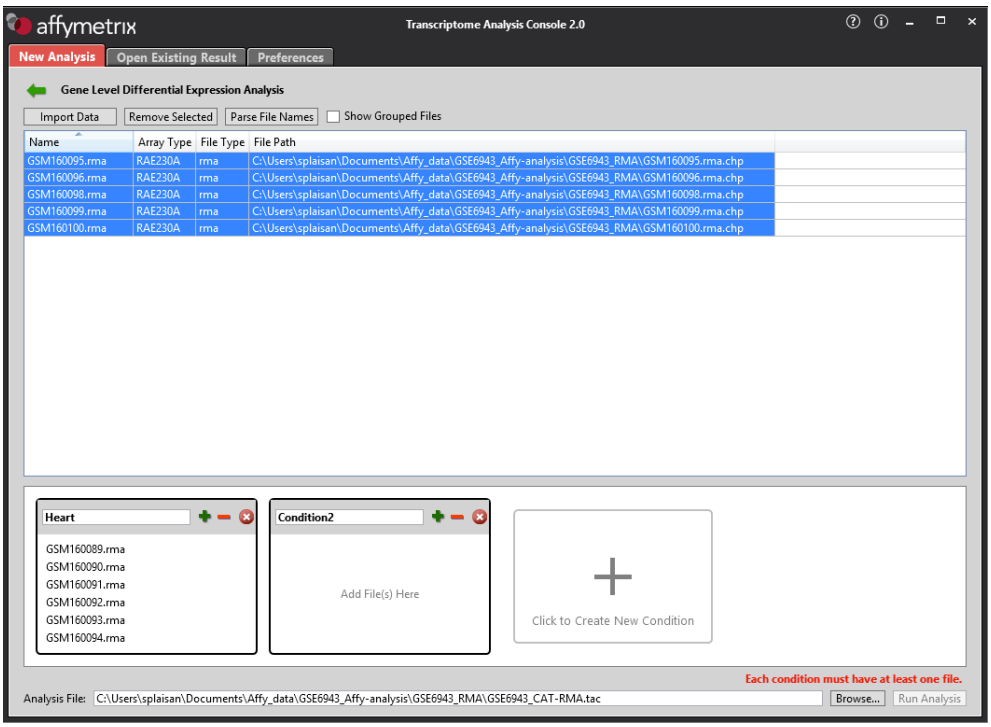
A number of QC plots can be generated using the right tools. The full QC report can then be saved as PDF file and is available both for the **RMA method (http://data.bits.vib.be/pub/trainingen/AffyECTAC2014/GSE6943_EC-rma.qc.PDF)**, **MAS5 method (http://data.bits.vib.be/pub/trainingen/AffyECTAC2014/GSE6943_EC-mas5.qc.PDF)**, and **PLIER method (http://data.bits.vib.be/pub/trainingen/AffyECTAC2014/GSE6943_EC-plier.qc.PDF)** on our server. Users are welcome to evaluate each QC plot by themselves using the data available on the server as input (see link at the bottom of this page)

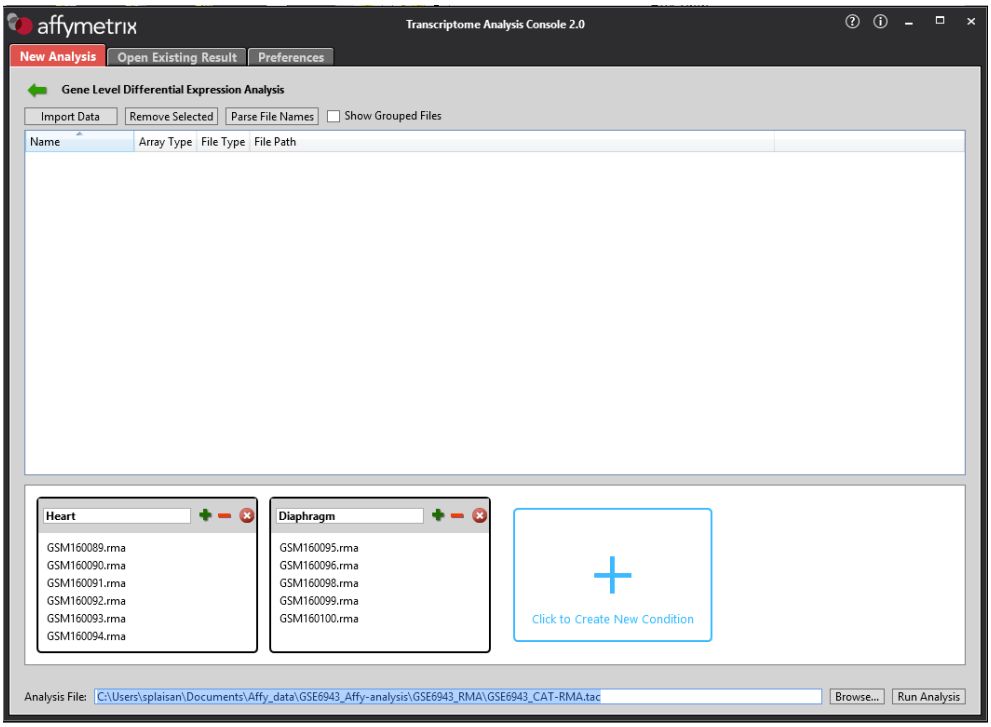
The Affymetrix Transcriptome Analysis Console (TAC)

Importing EC data and defining Groups



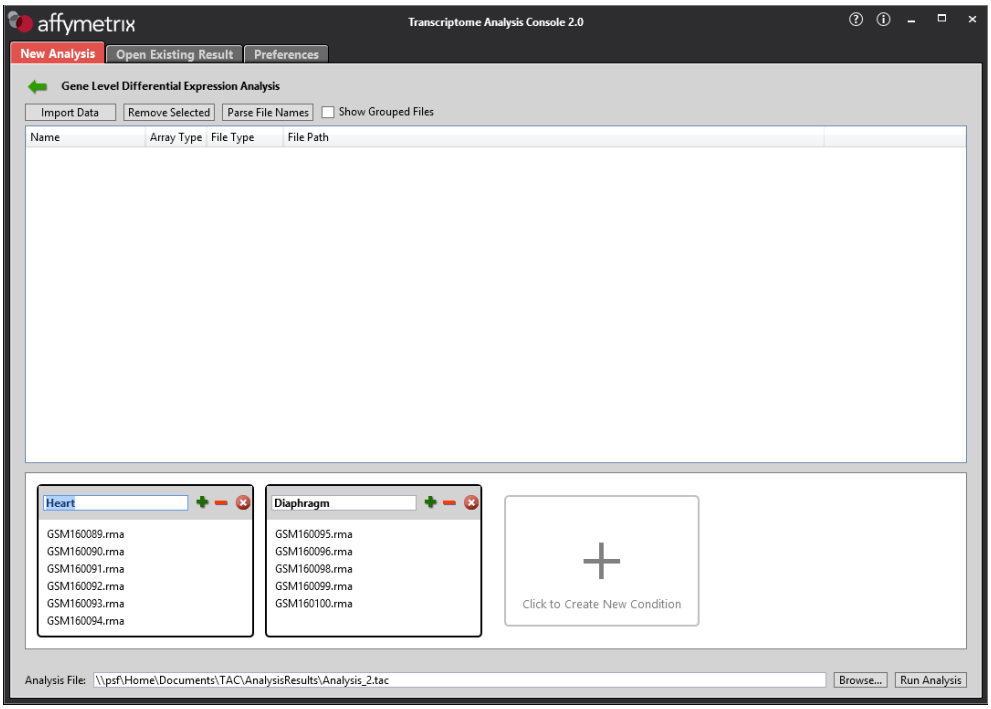
Each group is in turn defined by moving CHP files to the appropriate group window. This is done for 'Heart' and for 'Diaphragm' samples





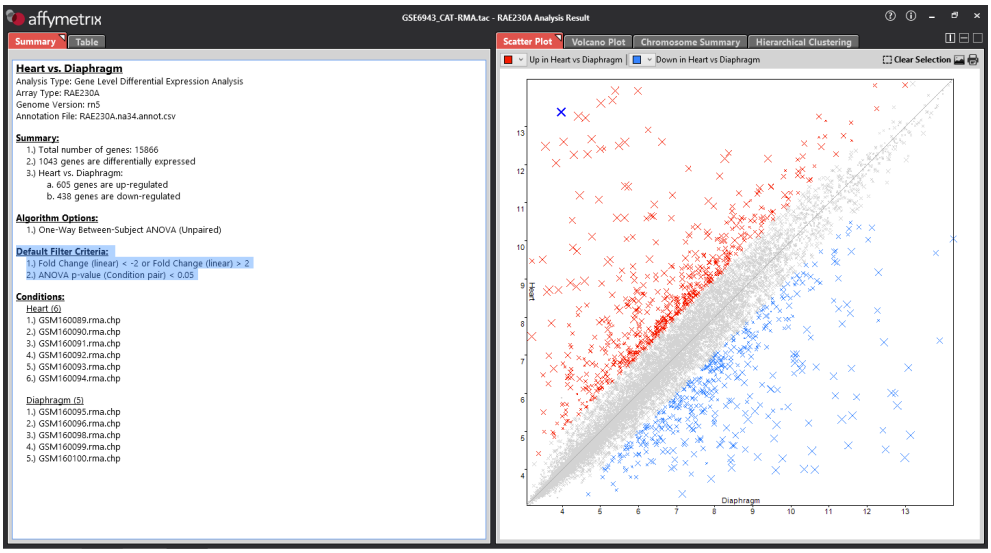
Computing 'gene-level' Differential expression

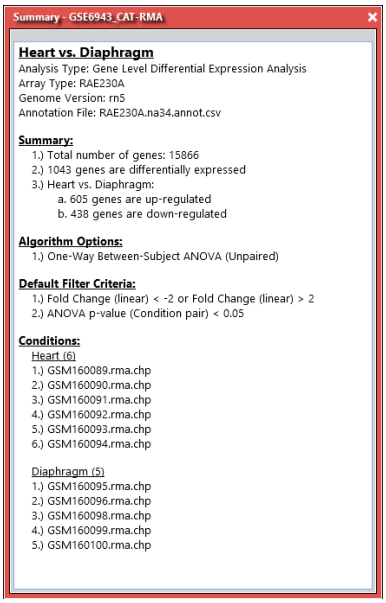
'Run analysis' is clicked to compute differential expression between the two groups



Other expression analyses can be performed when the probe type is compatible with transcript level analysis (discerning between alternative transcripts). However, this is not demonstrated here and we only provide the example of gene-level analysis.

The summary of a standard DE analysis is shown with counts for UR and DR genes under standard filtering values (more than two-fold difference between the groups and adjusted p-value < 0.05)

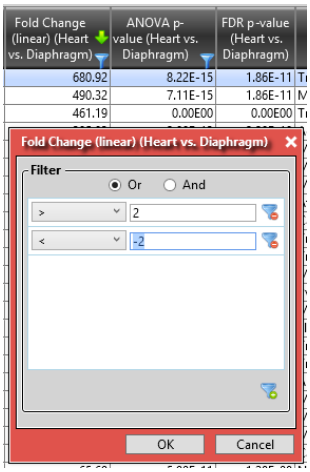




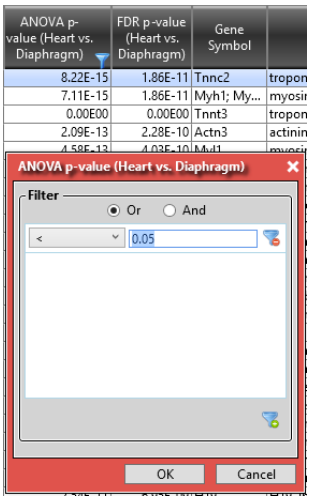
Adjusting Differential expression limits

The filtering values can be adapted by the user to restrain or increase the DE gene list and new plots generated.

Adjusting the differential expression limit

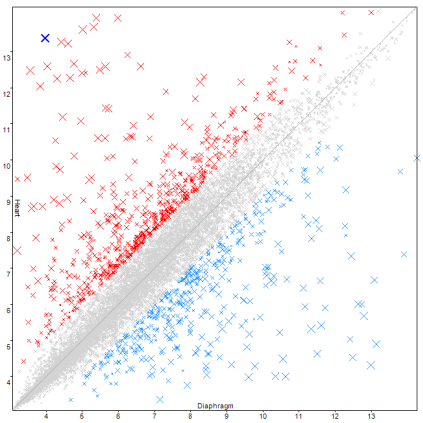


Adjusting the limit for the adjusted p-value

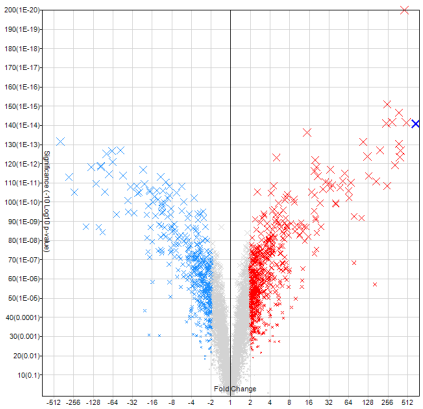


Plots based on the filtered differential expression table

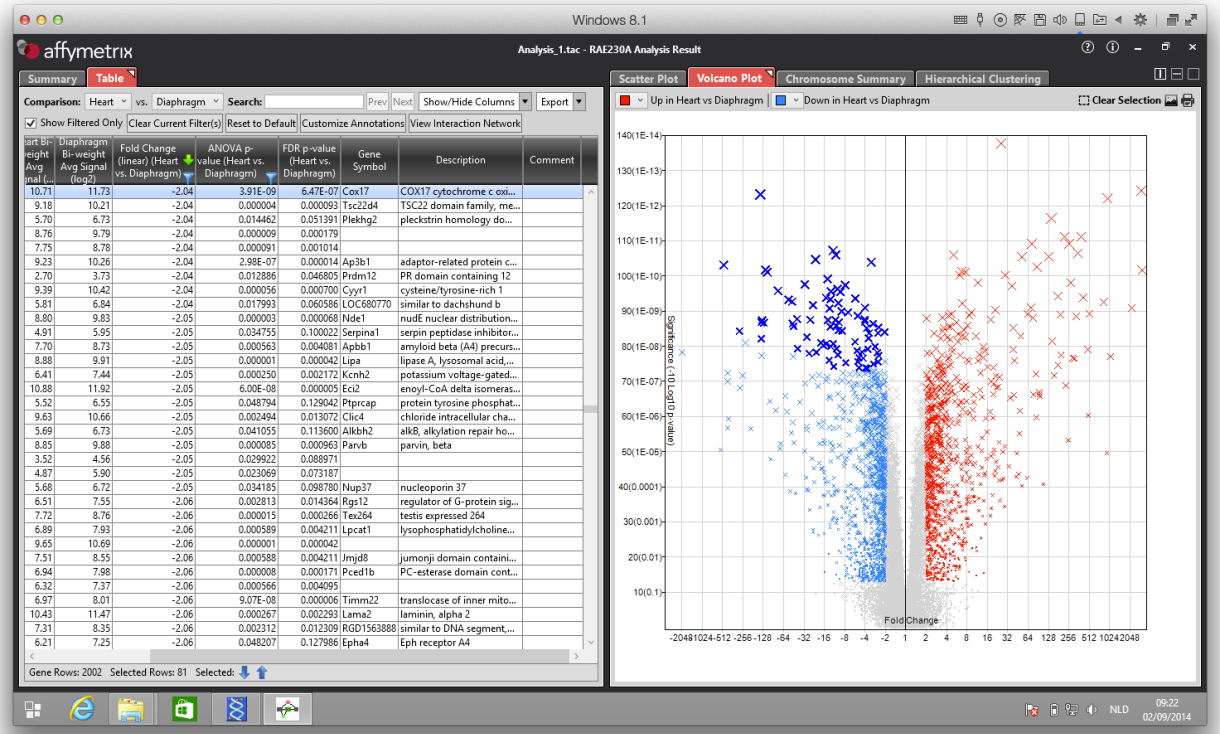
Additional graphs can be obtained to view the data from different angles. The scatter plot highlights potential differences between UR and DR genes between the groups. The graphs are interactive and the user can query the full data to find which probesets or genes are UR or DR using the mouse and selecting area around points.

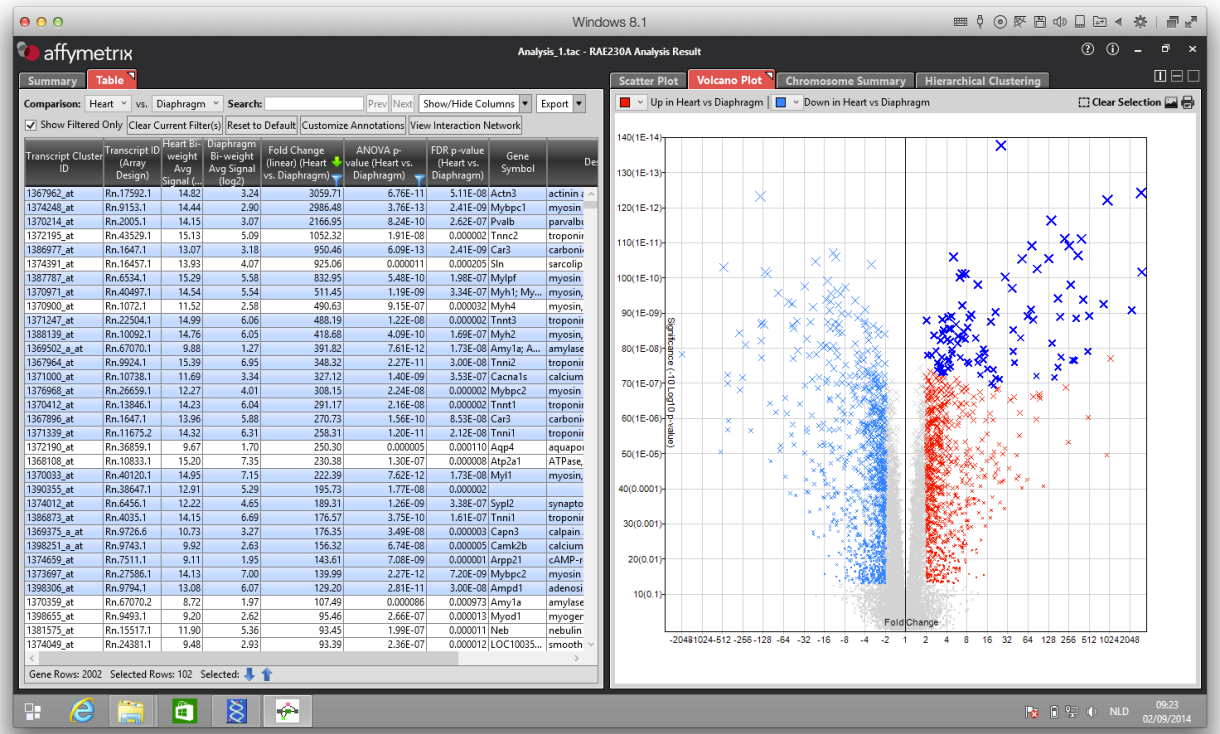


Volcano plots are very popular and show how confident the data is and how many genes show deviation from the steady state

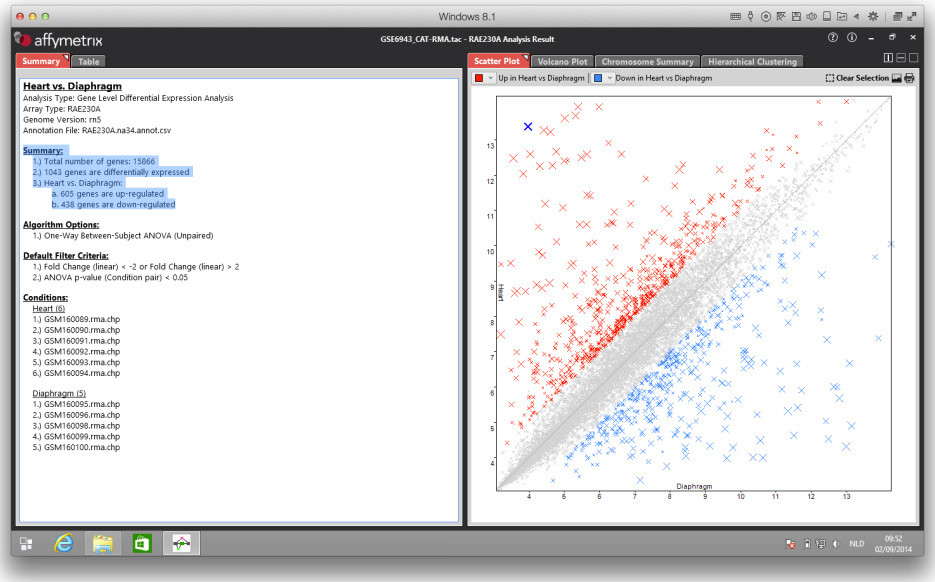


The interactive nature of the plot allows identifying outliers or significantly DE genes using the mouse.

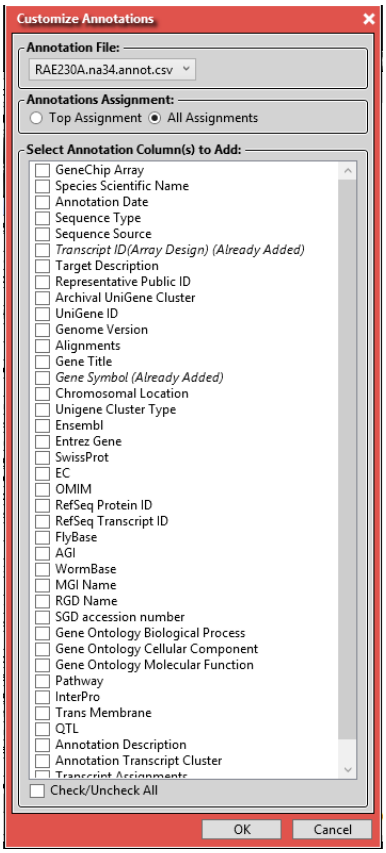

Windows 8.1
affymetrix
Analysis_1.tac - RAE230A Analysis Result
Summary
Table
Scatter Plot
Volcano Plot
Chromosome Summary
Hierarchical Clustering
Comparison: Heart vs. Diaphragm
Search:
Prev
Next
Show/Hide Columns
Export
Show Filtered Only
Clear Current Filter(s)
Reset to Default
Customize Annotations
View Interaction Network
Up in Heart vs Diaphragm
Down in Heart vs Diaphragm
Clear Selection
Transcript Cluster ID
Diaphragm Bi-weight Avg Signal (log2)
Fold Change (linear) (Heart vs. Diaphragm)
ANOVA p-value (Heart vs. Diaphragm)
FDR p-value (Heart vs. Diaphragm)
Gene Symbol
Description
Comment
10.71 11.73 -2.04 3.91E-09 6.47E-07 Cox17 COX17 cytochrome c oxi...
9.18 10.21 -2.04 0.000004 0.000093 Tsc22d4 TSC22 domain family, me...
5.70 6.73 -2.04 0.014462 0.051391 Plekhg2 pleckstrin homology do...
8.76 9.79 -2.04 0.000009 0.000179
7.75 8.78 -2.04 0.000091 0.001014
9.23 10.26 -2.04 2.98E-07 0.000014 Ap3b1 adaptor-related protein c...
2.70 3.73 -2.04 0.012886 0.046805 Prdm12 PR domain containing 12
9.39 10.42 -2.04 0.000056 0.000700 Cyyr1 cysteine/tyrosine-rich 1
5.81 6.84 -2.04 0.017993 0.060586 LOC680770 similar to dachshund b
8.80 9.83 -2.05 0.000003 0.000068 Ndel1 nudE nuclear distribution...
4.91 5.95 -2.05 0.034755 0.100022 Serpina1 serpin peptidase inhibitor...
7.70 8.73 -2.05 0.000563 0.004081 Apbb1 amyloid beta (A4) precurs...
8.88 9.91 -2.05 0.000001 0.000042 Lipa lipase A, lysosomal acid,...
6.41 7.44 -2.05 0.000250 0.002172 Kcnh2 potassium voltage-gated...
10.88 11.92 -2.05 6.00E-08 0.000005 Eci2 enoyl-CoA delta isomeras...
5.52 6.55 -2.05 0.048794 0.129042 Ptprcap protein tyrosine phosphat...
9.63 10.66 -2.05 0.002494 0.013072 Clic4 chloride intracellular cha...
5.69 6.73 -2.05 0.041055 0.113600 Alkbh2 alkB, alkylation repair ho...
8.85 9.88 -2.05 0.000085 0.000963 Parvb parvin, beta
3.52 4.56 -2.05 0.029922 0.088971
4.87 5.90 -2.05 0.023069 0.073187
5.68 6.72 -2.05 0.034185 0.098780 Nup37 nucleoporin 37
6.51 7.55 -2.06 0.002813 0.014364 Rgs12 regulator of G-protein sig...
7.72 8.76 -2.06 0.000015 0.000266 Tex264 testis expressed 264
6.89 7.93 -2.06 0.000589 0.004211 Lpcat1 lysophosphatidylcholine...
9.65 10.69 -2.06 0.000001 0.000042
7.51 8.55 -2.06 0.000588 0.004211 Jmjd8 jumonji domain containi...
6.94 7.98 -2.06 0.000008 0.000171 Pced1b PC-esterase domain cont...
6.32 7.37 -2.06 0.000566 0.004095
6.97 8.01 -2.06 9.07E-08 0.000006 Timm22 translocase of inner mito...
10.43 11.47 -2.06 0.000267 0.002293 Lama2 laminin, alpha 2
7.31 8.35 -2.06 0.002312 0.012309 RGD1563888 similar to DNA segment,...
6.21 7.25 -2.06 0.048207 0.127986 Epha4 Eph receptor A4
Gene Rows: 2002 Selected Rows: 81 Selected:
Significance (-1*Log10 p-value)
Fold Change
-2048 1024 -512 -256 -128 -64 -32 -16 -8 -4 -2 1 2 4 8 16 32 64 128 256 512 1024 2048
140(1E-14)
130(1E-13)
120(1E-12)
110(1E-11)
100(1E-10)
90(1E-09)
80(1E-08)
70(1E-07)
60(1E-06)
50(1E-05)
40(0.0001)
30(0.001)
20(0.01)
10(0.1)
NLD
09:22
02/09/2014



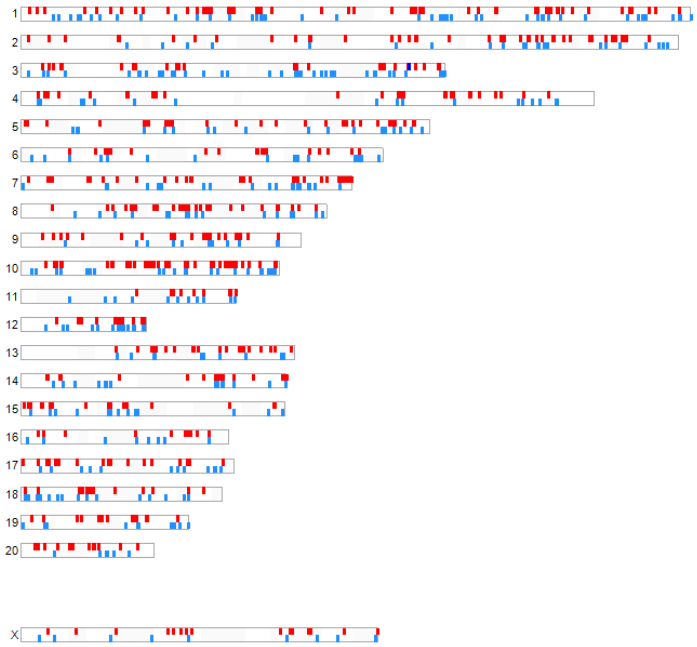
The count of UR and DR genes is reported in the summary page



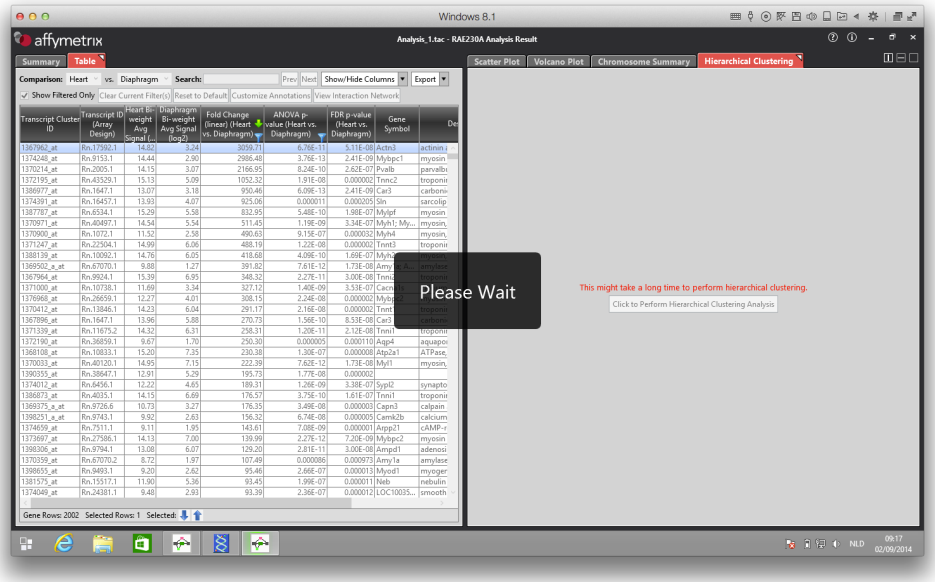
Additional annotations can be added using the dedicated menu

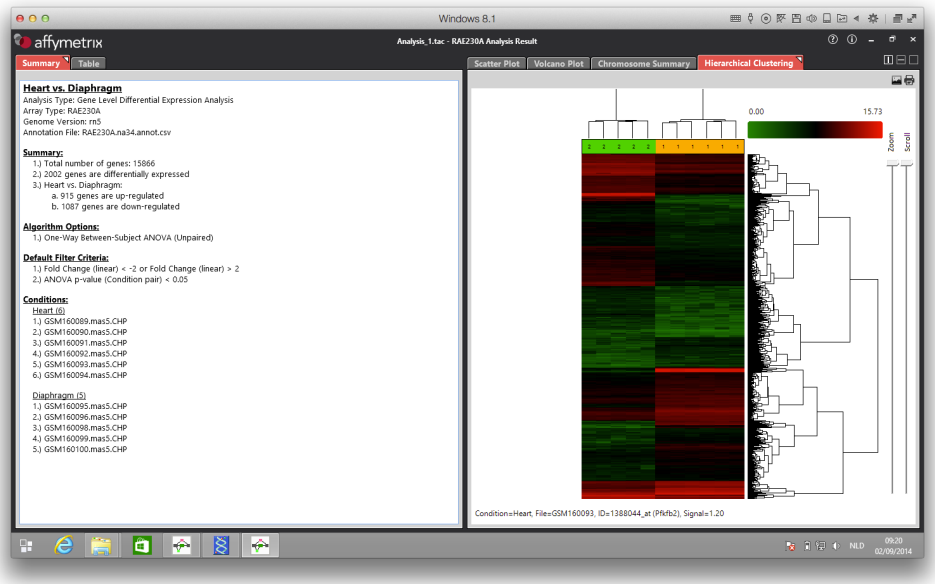


A plot of differential expression per chromosome may highlight local regulatory biases (hot spot loci)



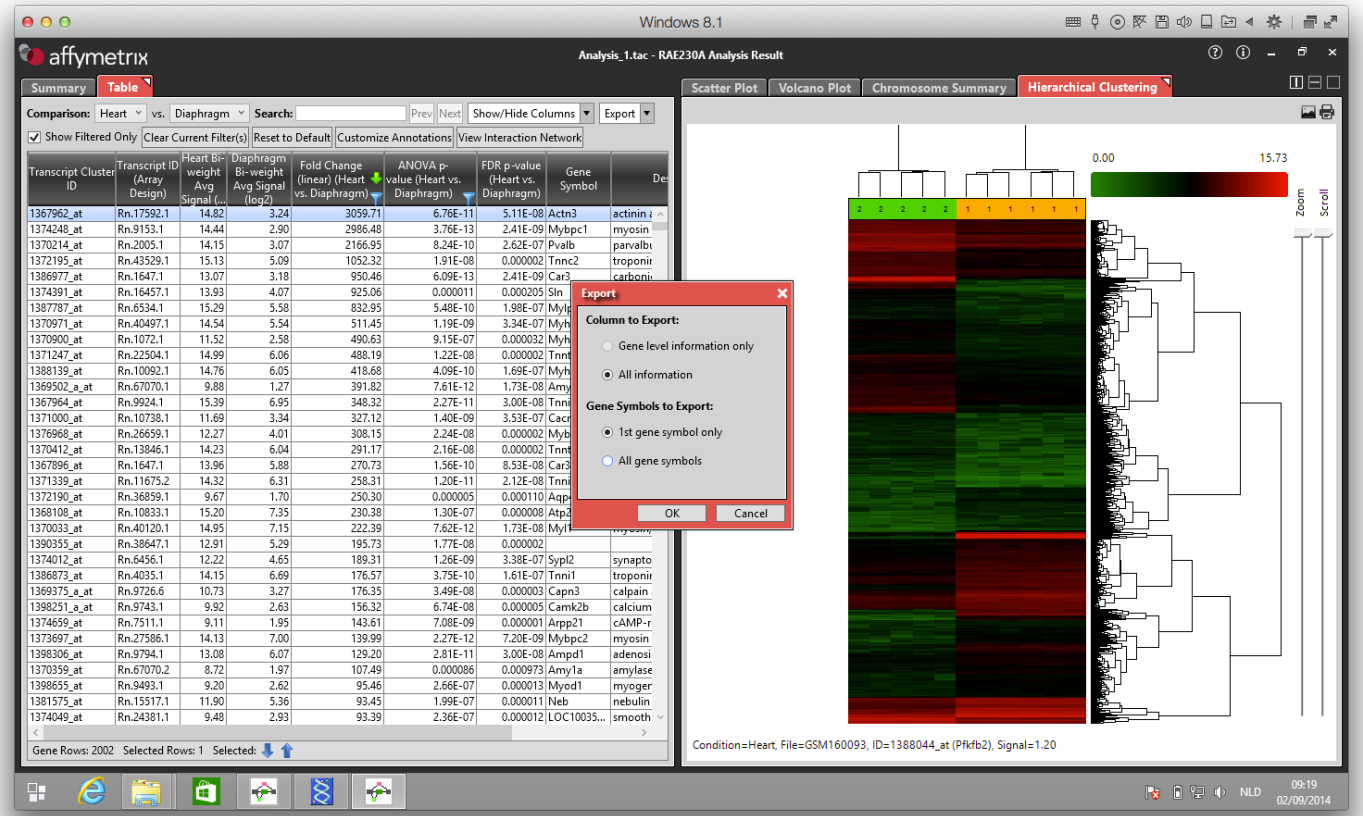
Heatmaps can be generated that show genes with similar pattern of variation across samples



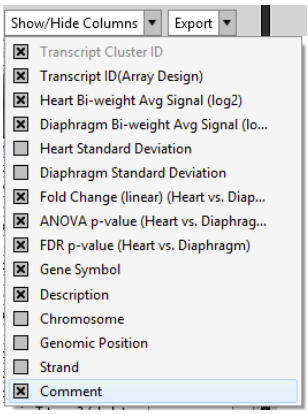


Exporting results

The tabular results can finally be exported to local file(s) for further use (IPA, ...)



Additional columns can be added to the table if the user needs them



After download to 'txt' files, results can easily be converted and filtered in the Excel spreadsheet editor

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Transcript Cluster ID	Transcript ID(Array Design)	Heart Bi-weight Avg Signal (log2)	Diaphragm Bi-weight Avg Signal (log2)	Heart Standard Deviation	Diaphragm Standard Deviation	Fold Change (linear) (Heart vs, Diaphrag	ANOVA p-value (Heart vs, Diaphrag	FDR p-value (Heart vs, Diaphragm)	Gene Symbol	Description	Chromosome	Genomic Position
2	1368093_at	Rn.54399.1	4,31	12,98	0,23	0,12	-407,47	6,94E-14	1,14E-10	Myh6	myosin, heavy chain 6, cardiac muscle, alpha	chr15	37492581-37516282
5	1367665_at	Rn.3789.1	4,9	13,14	0,29	0,28	-301,97	5,01E-12	2,21E-09	Ankrd1	ankyrin repeat domain 1	chr1	262038143-2620467
8	1367664_at	Rn.3789.1	4,52	12,48	0,43	0,19	-249,07	3,07E-11	7,60E-09	Ankrd1	ankyrin repeat domain 1	chr1	262038143-2620467
9	1367592_at	Rn.9965.1	5,03	12,41	0,65	0,24	-166,63	1,89E-09	2,07E-07	Tnnt2	troponin T type 2 (cardiac)	chr13	57716336-57729185
10	1388876_at	Rn.1192.1	5,68	12,79	0,24	0,18	-138,33	1,52E-12	9,54E-10	Pln	phospholamban	chr20	36399680-36400626
11	1387049_at	Rn.54399.1	4,63	11,5	0,17	0,35	-116,45	1,10E-11	4,03E-09	Myh6	myosin, heavy chain 6, cardiac muscle, alpha	chr15	37492581-37516282
16	1369313_at	Rn.3849.1	6	12,73	0,62	0,1	-105,96	1,97E-09	2,09E-07	Fhl2	four and a half LIM domains 2	chr9	49591626-49620648
17	1367616_at	Rn.3835.1	4	10,62	0,14	0,26	-98,18	1,44E-12	9,53E-10	Nppb	natriuretic peptide B	chr5	168454272-1684556
18	1388597_at	Rn.28286.1	6,01	12,63	0,21	0,19	-98,15	1,26E-12	8,71E-10	Mybpc3	myosin binding protein C, cardiac	chr3	86666789-86667480
19	1386931_at	Rn.64141.1	7,38	13,9	0,63	0,14	-92,07	3,64E-09	3,38E-07	Tnni3	troponin I type 3 (cardiac)	chr1	75665118-75668803
21	1370157_at	Rn.9740.1	6,67	13,08	0,35	0,16	-85,15	2,96E-11	7,60E-09	Pln	phospholamban	chr20	36390885-36399430
22	1389727_at	Rn.18919.1	4,01	10,34	0,16	0,18	-80,08	3,05E-13	3,02E-10	LOC100911101; Lrrc10	leucine-rich repeat-containing protein 10-like; leucine-rich repeat-co	chr7	60099897-60100444
26	1374816_at	Rn.2307.1	5,41	11,61	0,18	0,25	-73,37	3,43E-12	1,76E-09	Gcom1	GRINL1A complex locus 1	chr8	75066585-75067085
30	1373987_at	Rn.9940.3	4,49	10,55	0,12	0,18	-66,56	2,16E-13	2,28E-10	Kcnip2; LOC100911951	Kv channel-interacting protein 2; Kv channel-interacting protein 2-lik	chr1	270360899-2703615
31	1384202_at	Rn.14758.1	5,31	11,33	0,16	0,19	-64,99	8,09E-13	6,11E-10	Tesc	tescalcin	chr12	45972889-45978551
33	1371951_at	Rn.3849.2	5,9	11,71	0,44	0,11	-56,08	4,33E-10	6,42E-08	Fhl2	four and a half LIM domains 2	chr9	49591196-49591681
35	1374672_at	Rn.3434.1	3,98	9,6	0,08	0,19	-49,27	1,98E-13	2,28E-10	Tnni3k	TNNI3 interacting kinase	chr2	279730852-2797313
38	1389411_at	Rn.19666.1	4,29	9,78	0,18	0,2	-45,17	4,10E-12	2,03E-09			chr3	120021190-1200220
44	1371677_at	Rn.3817.1	5,22	10,46	0,3	0,16	-37,91	8,22E-11	1,65E-08	Spink8	serine peptidase inhibitor, Kazal type 8	chr8	117258777-1172703
47	1367564_at	Rn.2004.1	5,94	11,16	0,18	0,4	-37,24	3,15E-10	5,05E-08	Nppa	natriuretic peptide A	chr5	168466312-1684676
48	1388506_at	Rn.7293.1	7,34	12,44	0,26	0,13	-34,43	1,60E-11	4,98E-09	Dsp	desmoplakin	chr17	29201855-29202516
50	1367949_at	Rn.10015.1	6,65	11,58	0,22	0,35	-30,56	3,73E-10	5,79E-08	Penk	proenkephalin	chr5	21834402-21839358
51	1370773_a_at	Rn.54469.1	4,33	9,12	0,14	0,24	-27,53	1,44E-11	4,75E-09	Kcnip2; LOC100911951	Kv channel-interacting protein 2; Kv channel-interacting protein 2-lik	chr1	270338004-2703602
52	1372539_at	Rn.3291.1	4,81	9,55	0,12	0,29	-26,69	3,45E-11	8,16E-09			chr14	17884583-17886375
53	1370229_at	Rn.81250.1	6,45	11,11	0,27	0,04	-25,23	3,04E-11	7,60E-09	Ndrg4	NDRG family member 4	chr19	9751516-9762475
56	1389532_at	Rn.7963.1	5,17	9,62	0,11	0,18	-21,89	2,05E-12	1,16E-09			chr17	85869857-85871678
57	1371566_at	Rn.15764.1	4,04	8,44	0,09	0,23	-21,02	1,12E-11	4,03E-09	Fbxl22	F-box and leucine-rich repeat protein 22	chr8	71869413-71872379
59	1398243_at	Rn.11345.1	7,91	12,29	1,24	0,1	-20,71	0,000044	0,000457	Csrp3	cysteine and glycine-rich protein 3 (cardiac LIM protein)	chr1	105206458-1052171
60	1370061_at	Rn.3788.1	5,64	9,92	0,13	0,34	-19,45	4,51E-10	6,50E-08	Rab3b	RAB3B, member RAS oncogene family	chr5	132356507-1324088

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Transcript Cluster ID	Transcript ID(Array Design)	Heart Bi-weight Avg Signal (log2)	Diaphragm Bi-weight Avg Signal (log2)	Heart Standard Deviation	Diaphragm Standard Deviation	Fold Change (linear) (Heart vs, Diaphragm)	ANOVA p-value (Heart vs, Diaphragm)	FDR p-value (Heart vs, Diaphragm)	Gene Symbol	Description	Chromosome	Genomic Position
2	1372195_at	Rn.43529.1	13,38	3,97	0,14	0,19	680,92	8,22E-15	1,86E-11	Tnnc2	troponin C type 2 (fast)	chr3	167429205-1674302
5	1370971_at	Rn.40497.1	12,48	3,55	0,17	0,13	490,32	7,11E-15	1,86E-11	Myh1; Myh2; Myh8	myosin, heavy chain 1, skeletal muscle, adult; myosin, heavy chain 2,	chr10	53514514-53517766
8	1371247_at	Rn.22504.1	13,27	4,42	0,08	0,1	461,19	0,00E+00	0,00E+00	Tnnt3	troponin T type 3 (skeletal, fast)	chr1	222573326-2225784
9	1367962_at	Rn.17592.1	13,23	4,6	0,18	0,26	395,93	2,09E-13	2,28E-10	Actn3	actinin alpha 3	chr1	227051529-2270674
10	1370033_at	Rn.40120.1	13,61	5,01	0,11	0,33	387,67	4,58E-13	4,03E-10	Myl1	myosin, light chain 1	chr9	73364156-73369886
11	1388139_at	Rn.10092.1	12,6	4,03	0,19	0,2	379,82	9,19E-14	1,33E-10	Myh2	myosin, heavy chain 2, skeletal muscle, adult	chr10	53487290-53491320
16	1387787_at	Rn.6534.1	13,94	5,38	0,09	0,16	377,88	2,22E-15	1,17E-11	Mylpf	myosin light chain, phosphorylatable, fast skeletal muscle	chr1	205649549-2056523
17	1368108_at	Rn.10833.1	13,69	5,32	0,14	0,35	330,79	1,20E-12	8,63E-10	Atp2a1	ATPase, Ca++ transporting, cardiac muscle, fast twitch 1	chr1	204836392-2048541
18	1367896_at	Rn.1647.1	12,04	3,83	0,17	0,09	296,23	6,88E-15	1,86E-11	Car3	carbonic anhydrase 3	chr2	107900426-1079092
19	1370412_at	Rn.13846.1	12,26	4,3	0,1	0,12	249,05	7,77E-16	6,17E-12	Tnnt1	troponin T type 1 (skeletal, slow)	chr1	75652128-75660589
21	1367964_at	Rn.9924.1	13,93	5,98	0,11	0,47	247,62	1,42E-11	4,75E-09	Tnni2	troponin I type 2 (skeletal, fast)	chr1	222505097-2225070
22	1374248_at	Rn.9153.1	12,67	4,76	0,15	0,12	240,41	7,77E-15	1,86E-11	Mybpc1	myosin binding protein C, slow type	chr7	29192152-29199672
26	1370214_at	Rn.2005.1	12,27	4,66	0,22	0,14	194,36	1,97E-13	2,28E-10	Pvalb	parvalbumin	chr7	119420291-1194352
30	1373697_at	Rn.27586.1	12,41	4,99	0,15	0,38	170,81	8,29E-12	3,46E-09	Mybpc2	myosin binding protein C, fast-type	chr1	101572108-1015759
31	1374391_at	Rn.16457.1	12,42	5,09	0,24	1,49	160,99	0,000002	0,00004	Sln	sarcolipin	chr8	57016720-57017242
33	1386873_at	Rn.4035.1	12,5	5,48	0,15	0,32	129,87	4,26E-12	2,04E-09	Tnni1	troponin I type 1 (skeletal, slow)	chr13	57676633-57684216
35	1371339_at	Rn.11675.2	12,59	5,64	0,24	0,09	123,72	4,13E-13	3,86E-10	Tnni1	troponin I type 1 (skeletal, slow)	chr13	57683074-57685257
38	1398306_at	Rn.9794.1	11,19	4,46	0,19	0,07	106,44	7,16E-14	1,14E-10	Ampd1	adenosine monophosphate deaminase 1	chr2	224999610-2250204
44	1376227_at	Rn.41395.1	12,91	6,25	0,13	0,58	101,15	6,68E-10	8,82E-08	Myoz1	myozenin 1	chr15	8185059-8185524
47	1376968_at	Rn.26659.1	10,54	4,26	0,48	0,17	77,47	5,60E-10	7,80E-08	Mybpc2	myosin binding protein C, fast-type	chr1	101591233-1015937
48	1370900_at	Rn.1072.1	9,47	3,2	0,94	0,05	76,98	1,46E-07	0,000006	Myh4	myosin, heavy chain 4, skeletal muscle	chr10	53552135-53553903
50	1386977_at	Rn.1647.1	11,08	4,96	0,2	0,27	69,53	9,73E-12	3,77E-09	Car3	carbonic anhydrase 3	chr2	107900426-1079092
51	1381575_at	Rn.15517.1	9,51	3,48	0,32	0,23	65,69	5,90E-11	1,28E-08	Neb	nebulin	chr3	42756010-42756391
52	1371298_at	Rn.3968.1	12,6	6,61	0,23	0,3	63,38	2,54E-11	6,95E-09	H19	H19, imprinted maternally expressed transcript (non-protein coding)	chr1	222639223-2226401
53	1373873_at	Rn.14050.1	11,44	5,64	0,13	0,26	55,6	3,17E-12	1,68E-09			chr17	34950334-34951353
56	1390355_at	Rn.38647.1	11,42	5,73	0,1	0,3	51,53	8,00E-12	3,43E-09			chr1	89135621-89137093
57	1369706_at	Rn.24079.1	9,83	4,27	0,16	0,29	47,28	1,69E-11	5,15E-09	Cacng1	calcium channel, voltage-dependent, gamma subunit 1	chr10	95655312-95668315

Conclusion

The combination of the Affymetrix Expression and Transcription Analysis Consoles allows Windows-PC users without any knowledge of [R] to perform standard analysis of Affymetrix microarray data and obtain differential expression tables that can be used for downstream biological interpretation. Note that other more specific options and alternative analysis workflows are available with the same tools and that this tutorial is only an introduction with a selection of basic methods.

The main added value of these tools over [R] are the full range of QC plots generated and classically produced by bioinformatician experts as well as the very rapid processing of public Affymetrix CEL data (within minutes). We therefore recommend exploring the EC and TAC tools and associate them to IPA and other downstream tools allowing biological evaluation of public microarray data.

Youtube videos from the Affymetrix training team

Please follow the video webcasts below to get familiar with the Affymetrix **Expression Console** and **Transcription Analysis Console**

A series of YouTube videos can be found on the Affymetrix web site *[Expand]*

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑ http://www.affymetrix.com/estore/browse/level_seven_software_products_only.jsp?productId=131414&categoryId=35623&productName=Affymetrix%2526%2523174%253B-Expression-Console%2526%2523153%253B-Software#1_1
2. ↑ http://www.affymetrix.com/support/learning/training_tutorials/tac_ec/index.affx#1_2

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=Analyze_GEO_data_with_the_Affymetrix_software&oldid=11773"
Category: Howto

- This page was last modified on 19 October 2014, at 20:32.
- This page has been accessed 164 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

Normalize CEL files with RMAExpress

From BioWareWIKI

Convert CEL files to normalized text tables

Contents

- 1 Introduction
- 2 RMAeXpress run with the GSE6943 data
  - 2.1 Convert CEL data
  - 2.2 Plot normalized data
  - 2.3 Convert data with the Convertor tool
  - 2.4 Final result
- 3 download exercise files

Introduction

Many programs, like CLC require normalized microarray data as input and do not support the CEL format. RMA_eXpress and its companion convertor tool rapidely produce normalized and log-transformed data from a collection of CEL files and the corresponding CDF annotation database and without the need to install R and bioconductor packages.

RMA eXpress was cited in several publications (eg [1]) and is available for **Windows** AND for **mac OSX** from its developer site [2]. A manual is also available **here (http://rmaexpress.bmbolstad.com/RMAExpress_UsersGuide.pdf)**

The Affymetrix tools produce the same kind of data with more QC plots and are preferred if you want to have a close look to your data

RMAeXpress run with the GSE6943 data

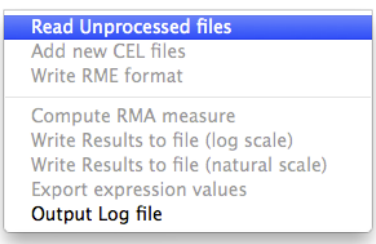
In order to perform this exercise, please first install the software from the following link: **http://rmaexpress.bmbolstad.com/**

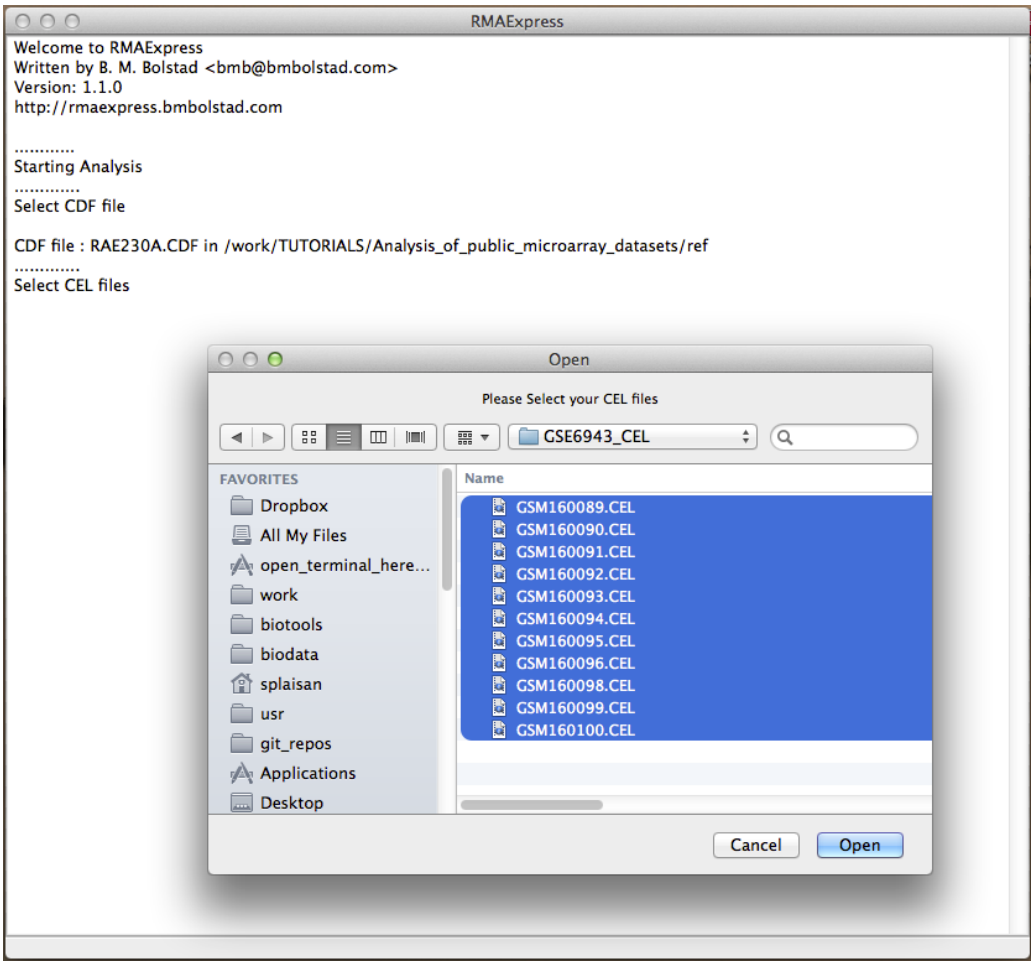
*You will need the rat **RAE230A.CDF** file accessible from the link at the bottom of this page*

Convert CEL data

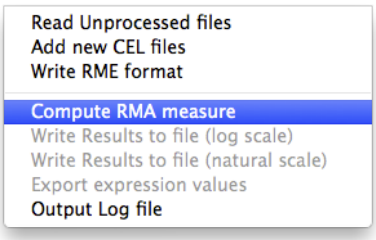
The user locates the CEL folder and the CDF file.

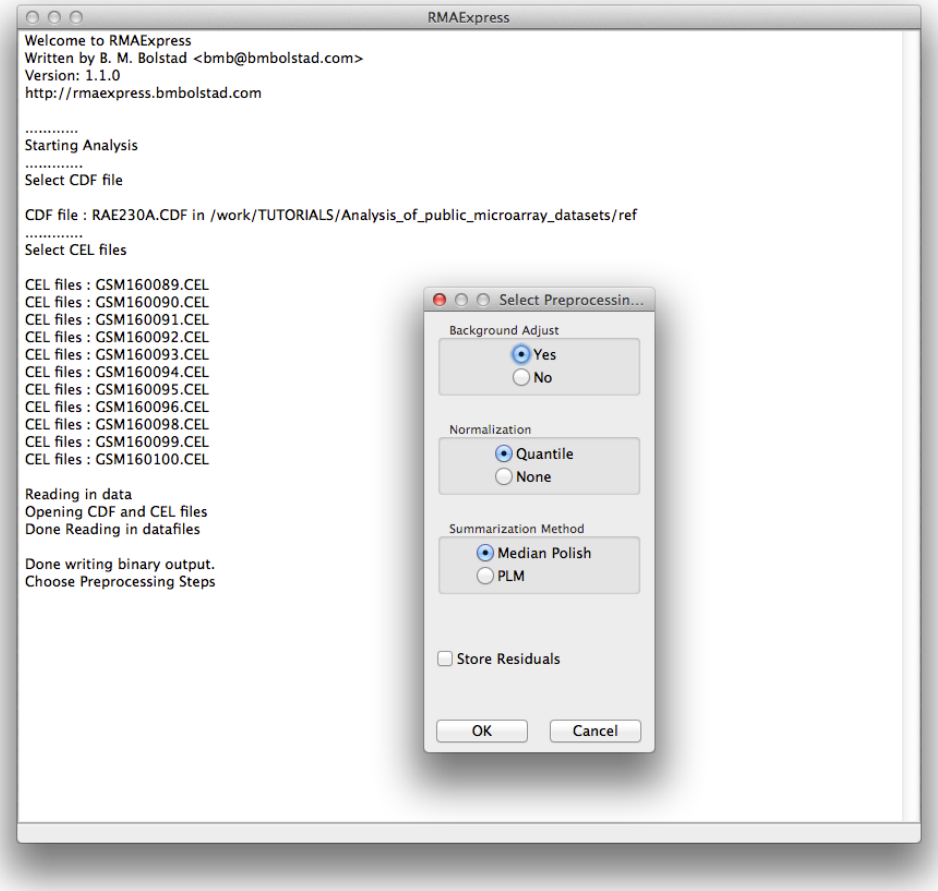
obviously all CEL files should come from the same CHIP



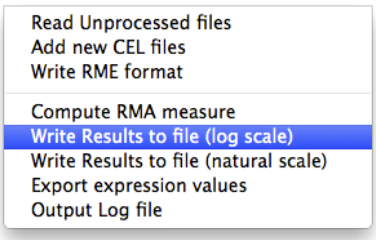


The RMA conversion is operated with user-selected options, leading to the activation of new menu items



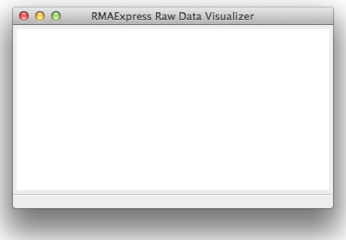


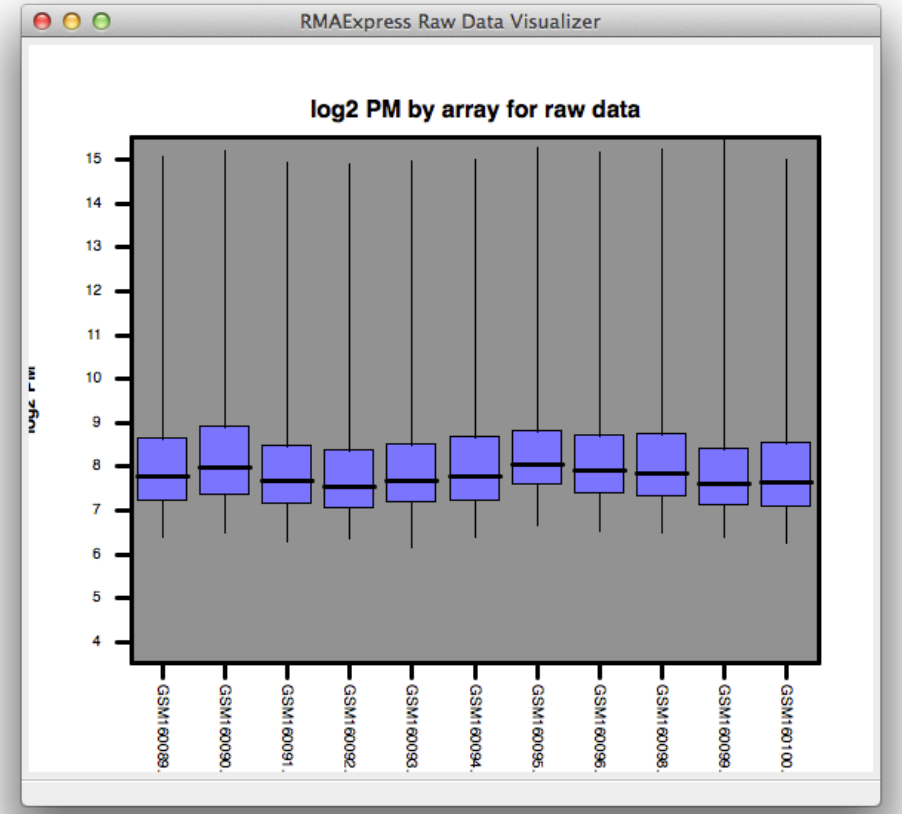
The log-tansformed results are saved to file for use with other programs

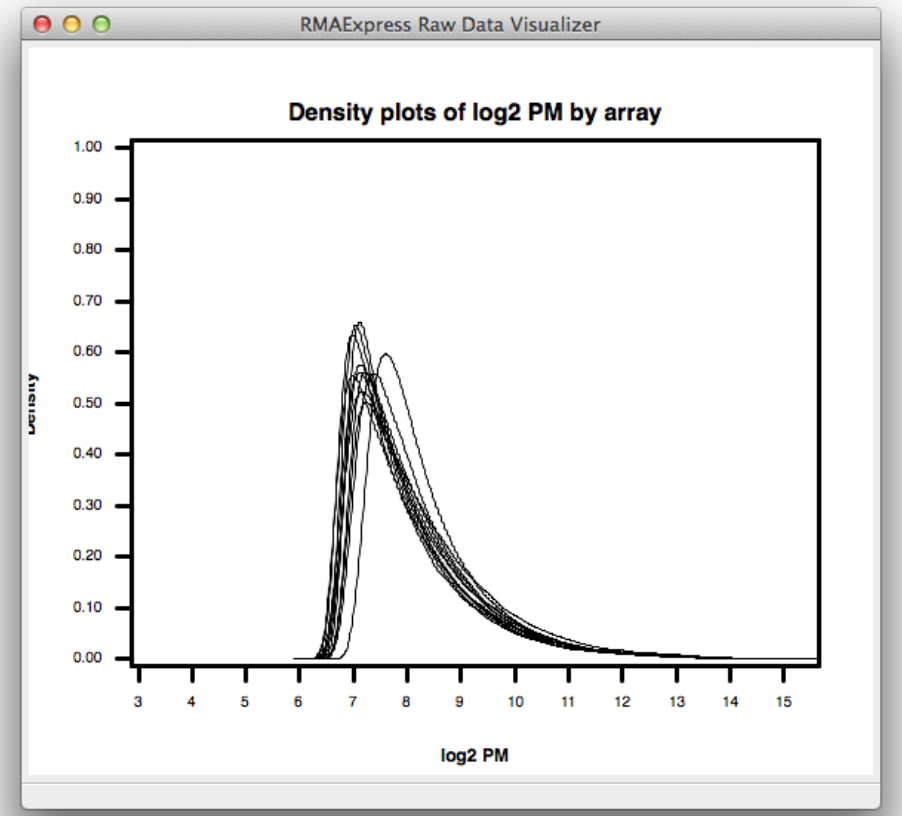


Plot normalized data

A number of QC plots are produced to inspect the normalized data

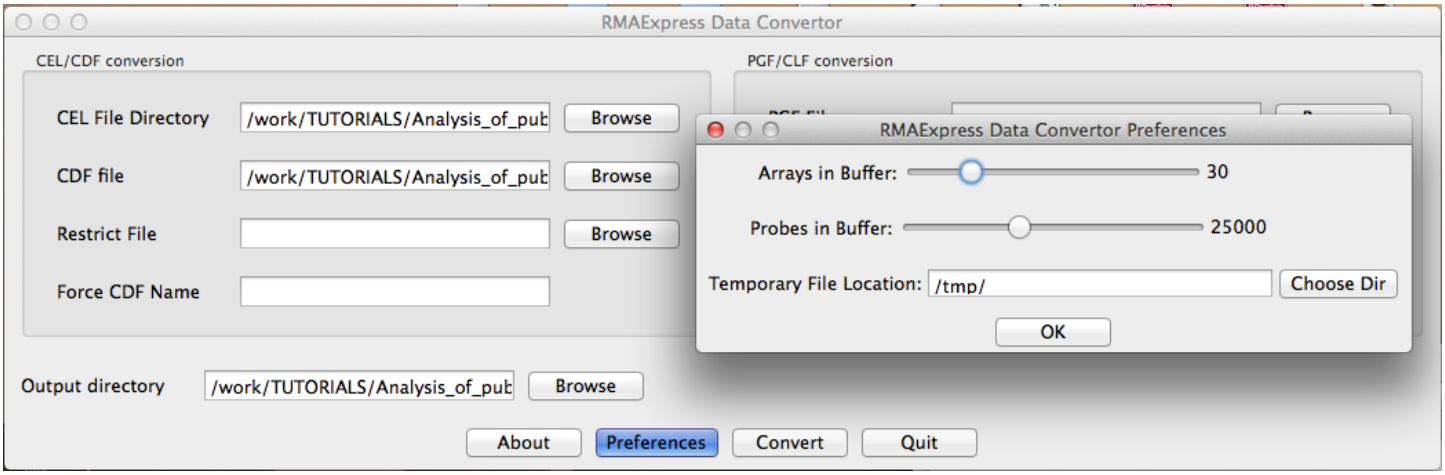






Convert data with the Convertor tool

The companion tool **RMA convertor** allows direct batch conversion without QC plots.



Final result

The resulting data for the GSE experiment is shown below (top 5 lines by first 5 columns for readability)

```
Probesets   GSM160089.CEL  GSM160090.CEL  GSM160091.CEL  GSM160092.CEL
1367452_at  9.642959       9.651789       9.328636       9.484756
1367453_at  10.155364      10.249902      10.186186      10.096563
1367454_at  8.739058       9.082530       9.071610       8.687629
1367455_at  10.476362      10.447124      10.571050      10.326480
```

download exercise files

Download exercise files here. *[Expand]*

References:

1. ↑

   Claudia Mimoso, Ding-Dar Lee, Jiri Zavadil, Marjana Tomic-Canic, Miroslav Blumenberg
   Analysis and meta-analysis of transcriptional profiling in human epidermis.
   Methods Mol. Biol.: 2014, 1195;61-97
   [PubMed:24297317] ##WORLDCAT## [DOI] (I p)

   Małgorzata Janas-Kozik, Urszula Mazurek, Irena Krupka-Matuszczyk, Małgorzata Stachowicz, Joanna Głogowska-Ligus, Tadeusz Wilczok
   The transcript expression profile of the leptin receptor-coding gene assayed with the oligonucleotide microarray technique--could this be an anorexia nervosa marker?
   Cell. Mol. Biol. Lett.: 2006, 11(1);62-9
   [PubMed:16847749] ##WORLDCAT## [DOI] (I p)

   Yoseph Barash, Elinor Dehan, Meir Krupsky, Wilbur Franklin, Marc Geraci, Nir Friedman, Naftali Kaminski
   Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays.
   Bioinformatics: 2004, 20(6);839-46
   [PubMed:14751998] ##WORLDCAT## [DOI] (P p)

2. ↑ http://rmaexpress.bmbolstad.com/

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=Normalize_CEL_files_with_RMAExpress&oldid=11803"
Category: Howto

- This page was last modified on 20 October 2014, at 10:08.
- This page has been accessed 20 times.

- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

Find Transcriptome Signatures with TranscriptomeBrowser

From BioWareWIKI

Search public GEO datasets, identify specific transcriptome signatures, and perform functional enrichment on the found sets

Contents

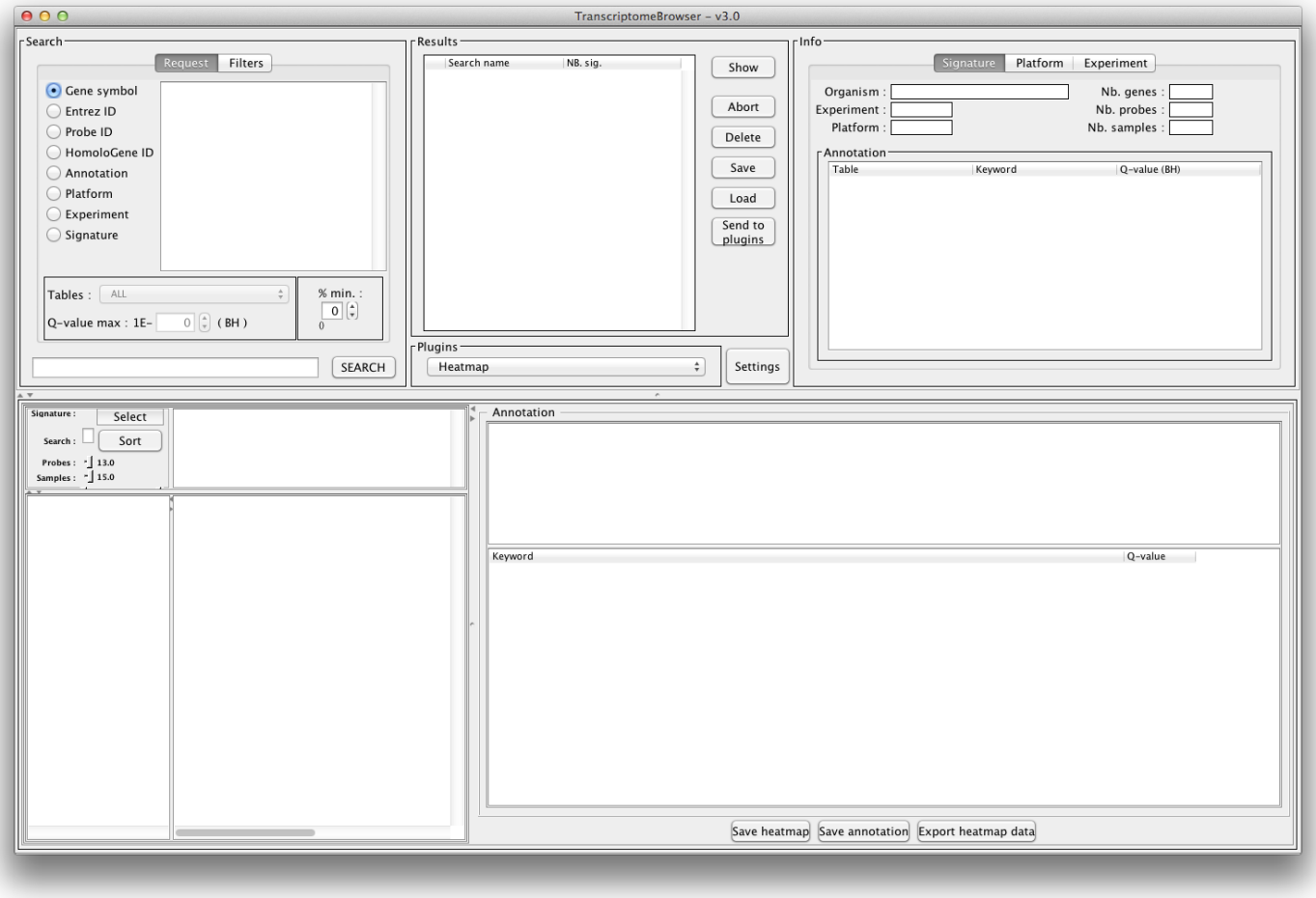
- 1 Introduction
- 2 A Walk-through example
  - 2.1 Load the GSE6943 dataset
  - 2.2 Find Transcriptome Signatures
  - 2.3 Show HeatMaps for each TS
- 3 Conclusion
  - 3.1 And there is even more for Geeks

Introduction

TranscriptomeBrowser (TBrowser)[1] host a large database of transcriptional signatures (TS) extracted from GEO public microarray repository [2]. TS have been produced using a new algorithm called "Density Based Filtering And Markov Clustering" (DBF-MCL). The current database contains about 30 000 TS derived from ~ 4 000 microarray datasets (~222 millions expression values). Each TS was tested for functional enrichment using annotation obtained from numerous ontologies or annotation databases (Gene Ontology, KEGG, BioCarta, Swiss-Prot, BBID, SMART, NIH Genetic Association DB, COG/KOG, TargetScan, PicTar, .TFBS_conserv.ed., MSigDB, GeneSigDB,...). TBrowser comes with a sophisticated search engine so that users can perform combined queries using boolean operators.

VERSION 3.0. TranscriptomeBrowser host a large database of transcriptional signatures (TS, n~40 000) extracted from Gene Expression Omnibus (~4 000 experiments) using the DBF-MCL algorithm. TBrowser comes with a sophisticated search engine so that users can search for the biological contexts in which several genes were concomitantly regulated. Several examples are provided below and in the article published in PLoSONE [1]. A video tutorial for **TranscriptomeBrowser** is available **here (https://www.youtube.com/watch?v=_bJMEPe5gHI)** and a second for the **InteractomeBrowser** plugin **here (https://www.youtube.com/watch?v=SxOBmCP1G1A)**. The full manual can be read **here (http://tagc.univ-mrs.fr/tbrowser/index2.php?option=com_content&task=view&id=19&pop=1&page=0&Itemid=23)**[3]

After installation (see online documentation) and startup, the main TBrowser interface awaits user input

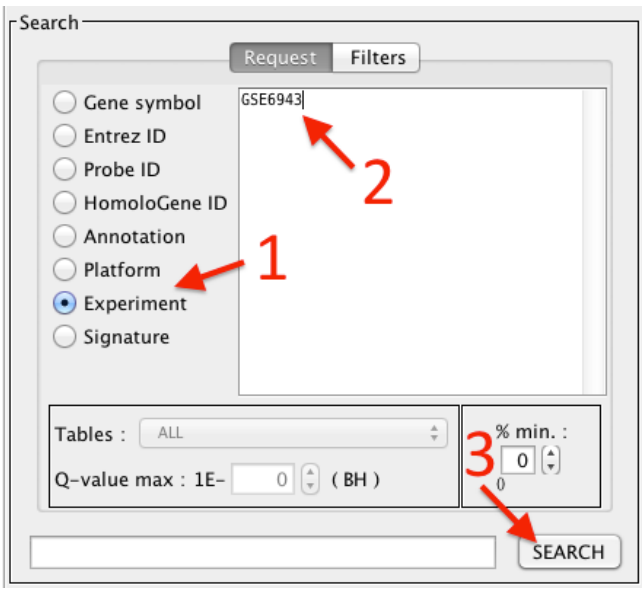


A Walk-through example

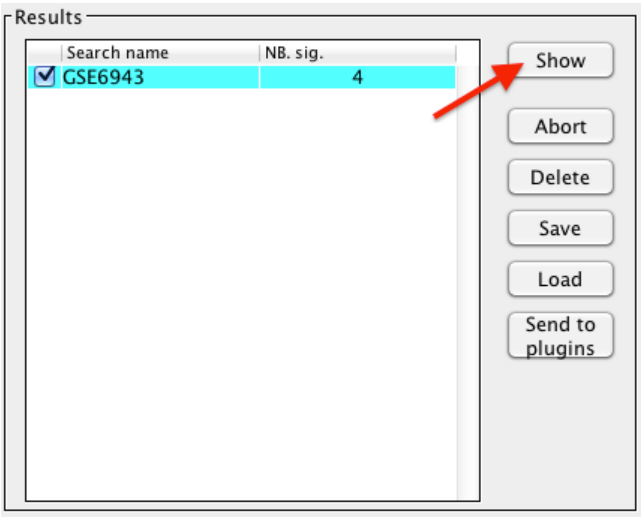
As shown above, different search angles can be used to populate the interface. Instead of searching for genes, we choose to use here the GSE6943 dataset used in other BITS training tutorial(s). Please refer to the **Hands-on_Analysis_of_public_microarray_datasets** page for more information about this dataset.

Load the GSE6943 dataset

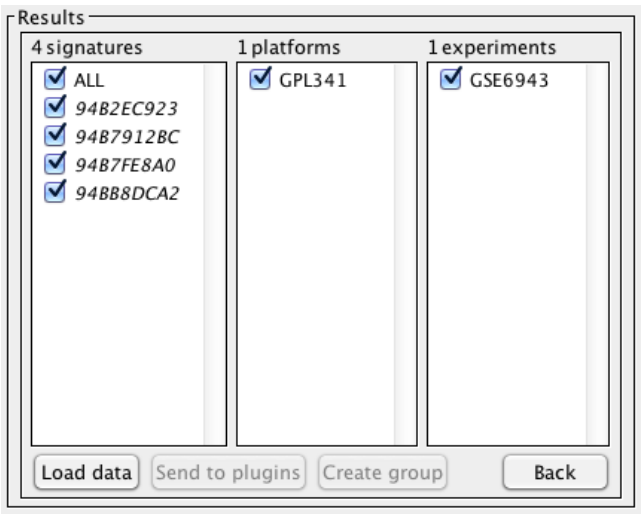
Change focus to search for 'Experiment' and type GSE6943 in the text field followed by a click on 'SEARCH'



Four enriched TS (*Transcriptome Signatures*) are found by the program



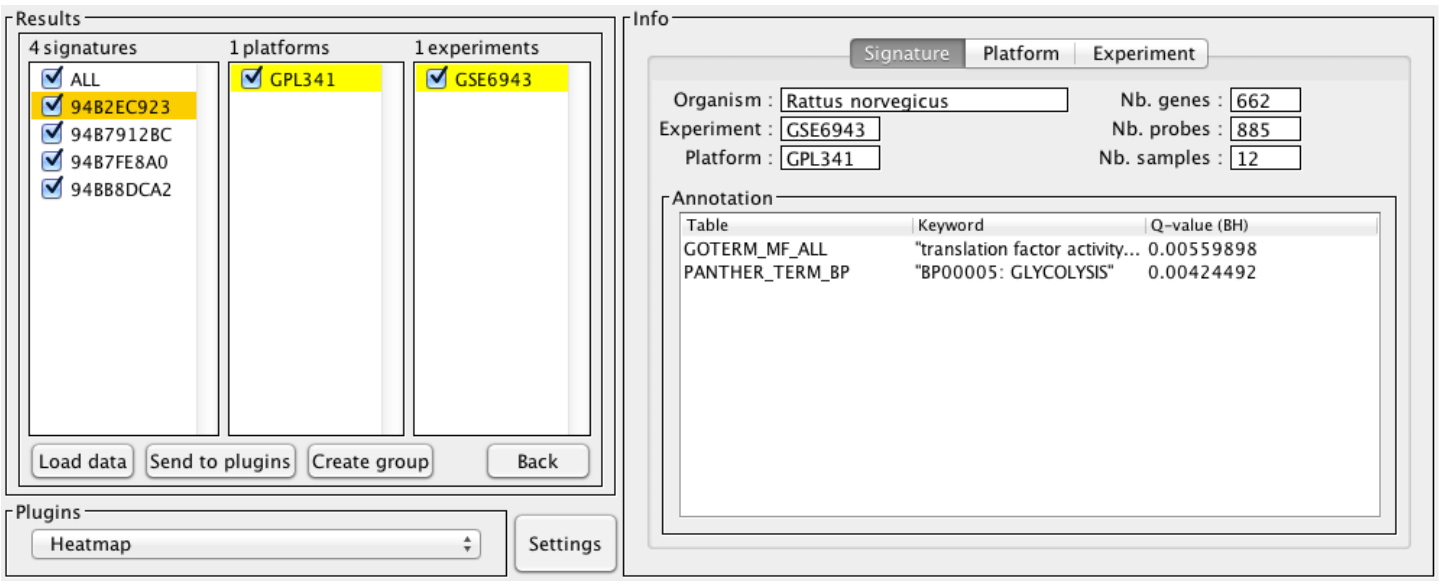
Several options are available here (please explore) and we choose to show the results to get details about the four hits



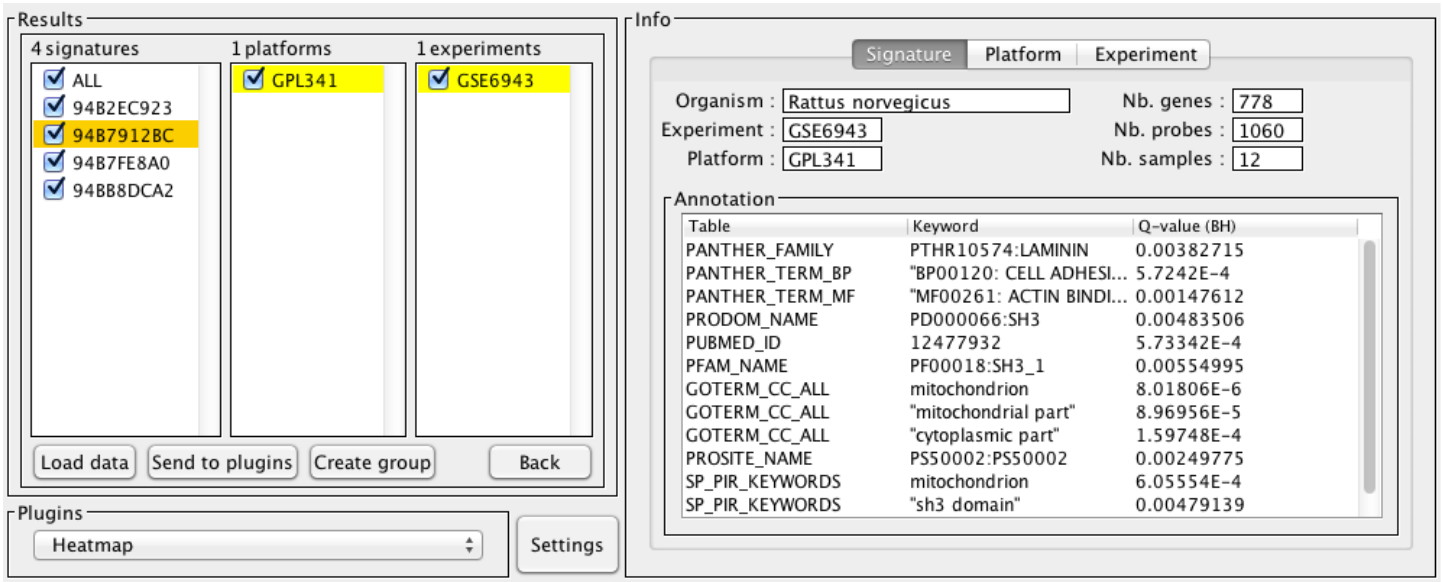
Find Transcriptome Signatures

Selecting each of the four hits and displaying the detailed statistical results on the top-right window

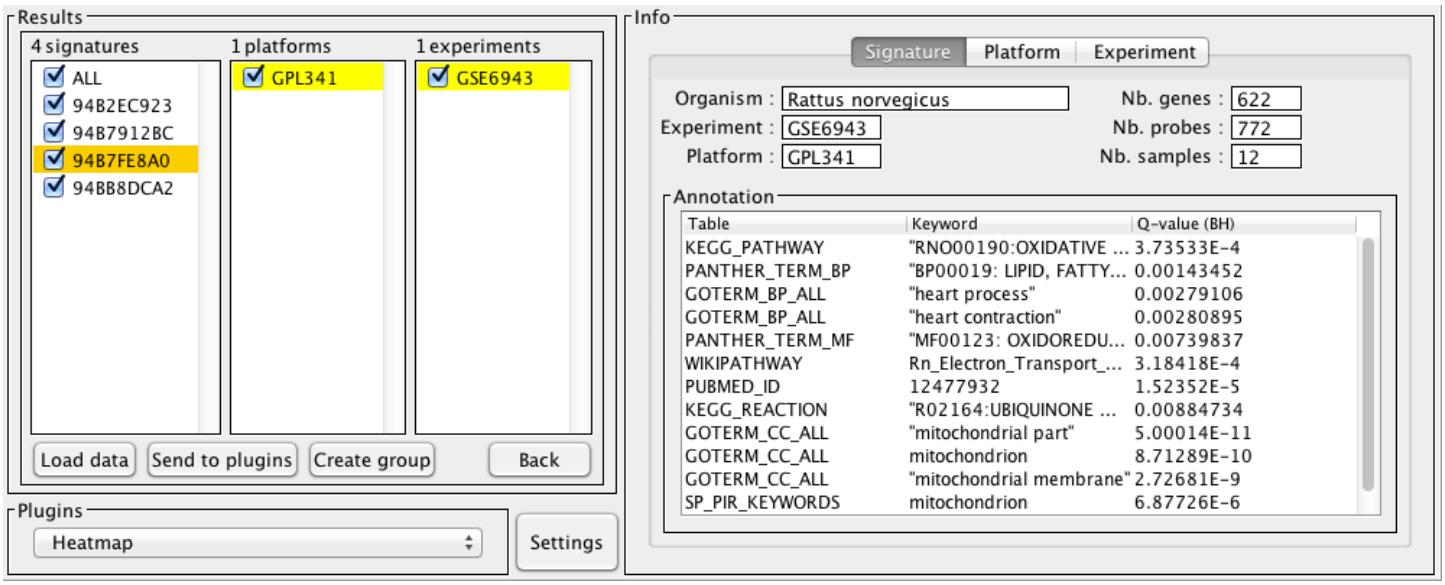
- first TS (662 genes) and **Glycolysis** as main aspect



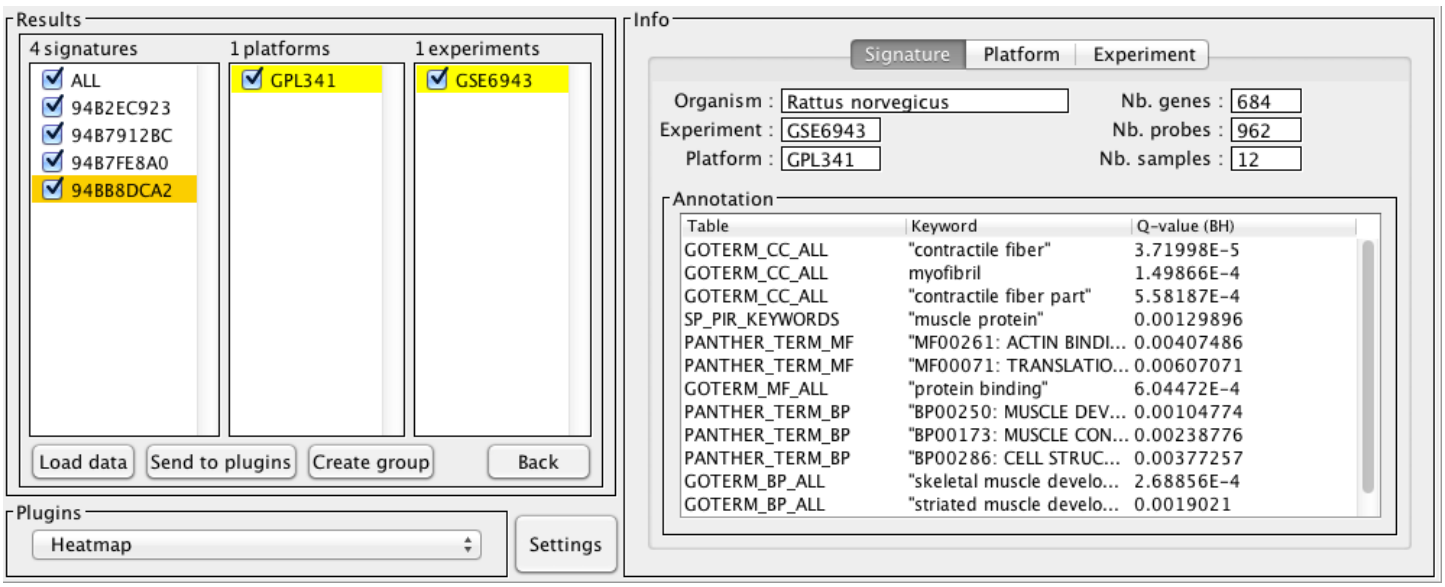
- second TS (778 genes) related to **mitochondrial functions**



- third TS (622) clearly associated with **heart and mitochondrial functions**

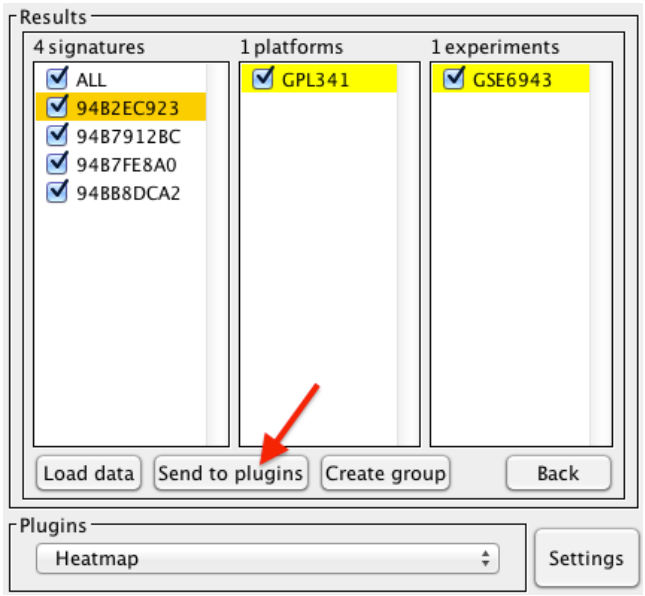


- last TS (684) apparently specific to the **cardiac muscle**

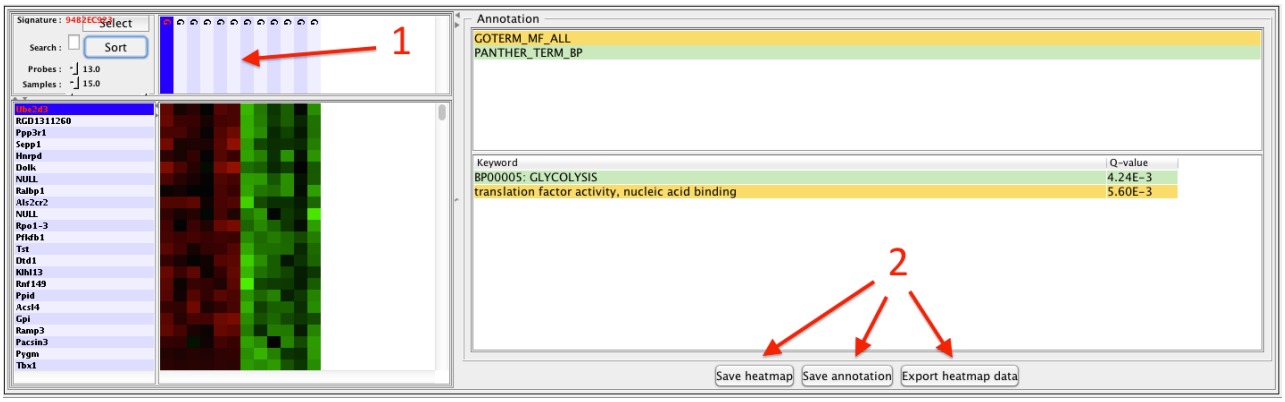


Show HeatMaps for each TS

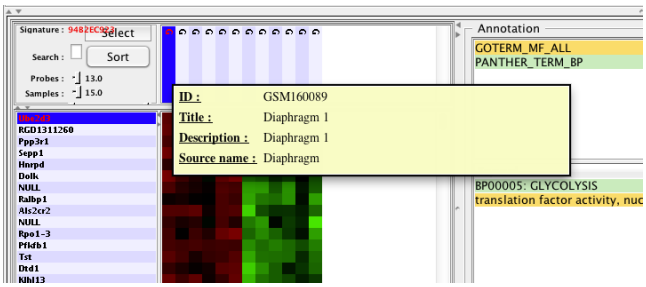
For each TS, we can visualize the actual expression data as a heatmap by using the corresponding plugin



- first heatmap (cut at ~20 genes)



the leftmost 6 samples are the **Diaphragm** replicates while the remaining ones are the **Heart** replicates

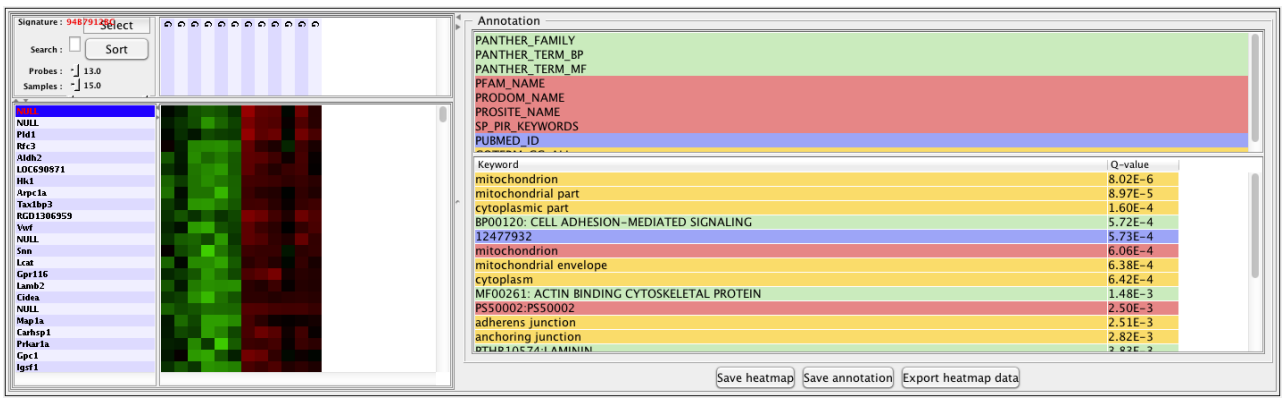


The full heatmap picture can be saved to file using buttons present at the bottom of the window

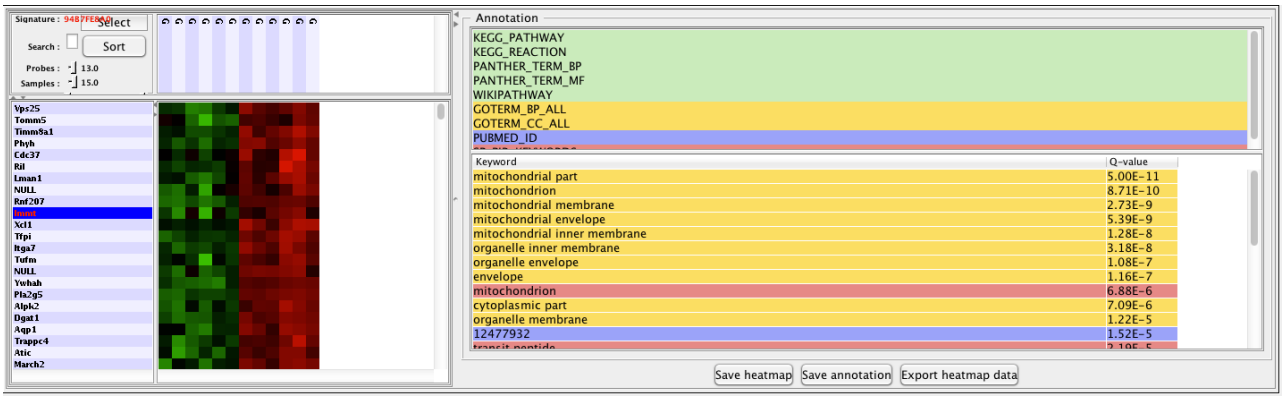
full heatmap#1 *[Expand]*

Similarly, text files can be saved with the **data used to plot the heatmaps (http://data.bits.vib.be/hidden/jhslbjcgnchjdgksqngcvgqdlsjcnv/TBrowser2014/09d_TS1-heatmap-data.txt)** and a text **table reporting enriched terms (http://data.bits.vib.be/hidden/jhslbjcgnchjdgksqngcvgqdlsjcnv/TBrowser2014/09e_TS1-terms.txt)**

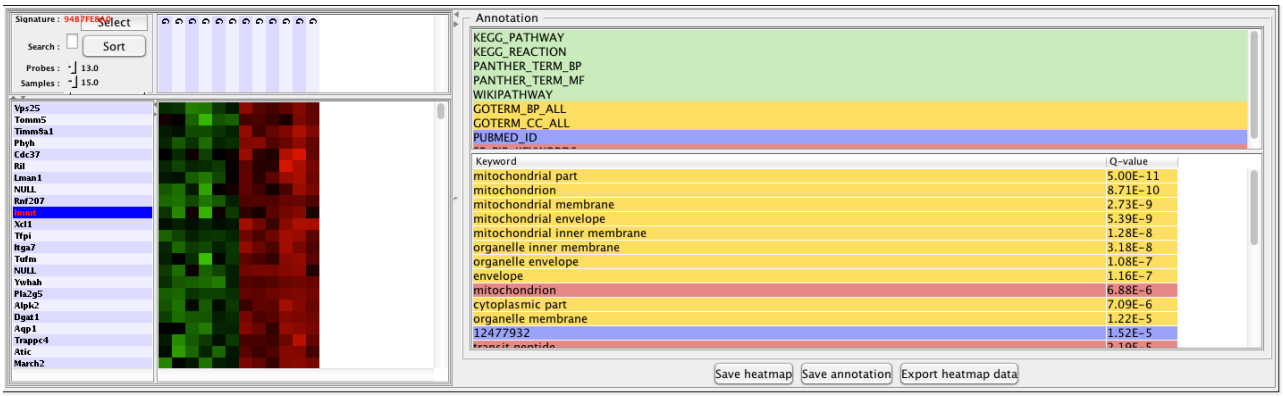
- second heatmap



- third heatmap



- fourth heatmap



Other buttons and tabs allow inspecting details of any particular view. Genes can be sorted alpha or by clustering order and a given gene can be searched using the search box

Conclusion

And there is even more for Geeks

The data mining package, **RTools4TB**, can perform calls to TranscriptomeBrowser web service and implements the DBF-MCL algorithm. The R package **RTools4TB (http://www.bioconductor.org/packages/2.5/bioc/html/RTools4TB.html)** is now part of the Bioconductor project.

References:

1. ↑ [1.0] [1.1] http://tagc.univ-mrs.fr/tbrowser/

   Cyrille Lepoivre, Aurélie Bergon, Fabrice Lopez, Narayanan B Perumal, Catherine Nguyen, Jean Imbert, Denis Puthier
   TranscriptomeBrowser 3.0: introducing a new compendium of molecular interactions and a new visualization tool for the study of gene regulatory networks.
   BMC Bioinformatics: 2012, 13;19
   [PubMed:22292669] ##WORLDCAT## [DOI] (I e)

   Fabrice Lopez, Julien Textoris, Aurélie Bergon, Gilles Didier, Elisabeth Remy, Samuel Granjeaud, Jean Imbert, Catherine Nguyen, Denis Puthier
   TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database.
   PLoS ONE: 2008, 3(12);e4001
   [PubMed:19104654] ##WORLDCAT## [DOI] (I p)

2. ↑ http://www.ncbi.nlm.nih.gov/gds/
3. ↑ http://tagc.univ-mrs.fr/tbrowser/index2.php?option=com_content&task=view&id=19&pop=1&page=0&Itemid=23

[Main_Page | Hands-on Analysis of public microarray datasets]

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=Find_Transcriptome_Signatures_with_TranscriptomeBrowser&oldid=11366"
Category: Howto

- This page was last modified on 8 September 2014, at 08:28.
- This page has been accessed 59 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.

PubMA Exercise.7

From BioWareWIKI

IPA analysis of the GEO2R DE table

The following content was obtained with the DE table generates by GEO2R. Very similar resuts are expected with the results from RobiNA or from the Affy console

Contents

- 1 Introduction
- 2 IPA Tutorial material
- 3 IPA analysis
  - 3.1 Upload data in IPA
  - 3.2 Start core analysis and set filter
  - 3.3 Review the obtained core results
- 4 download exercise files

Introduction

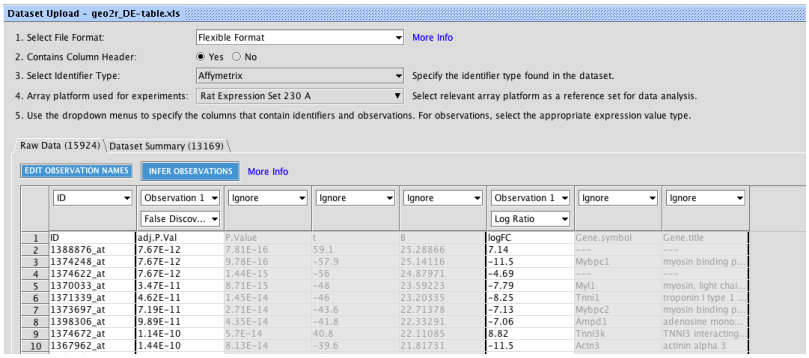
Ingenuity Pathway Analysis (IPA) is strongly advised for more advanced users/usage. You can use IPA on any Java-installed computer after asking for a personal account to **mailto:bits@vib.be** and login in **here (https://apps.ingenuity.com/ingsso/login?service=https%3A%2F%2Fanalysis.ingenuity.com%2Fpa%2Fj_spring_cas_security_check&originalUrl=https%3A%2F%2Fanalysis.ingenuity.com%2Fpa%3Futm_source%3DIngenuity%26utm_medium%3**. **Please keep in mind that IPA is only meant for human/mouse/rat data**.

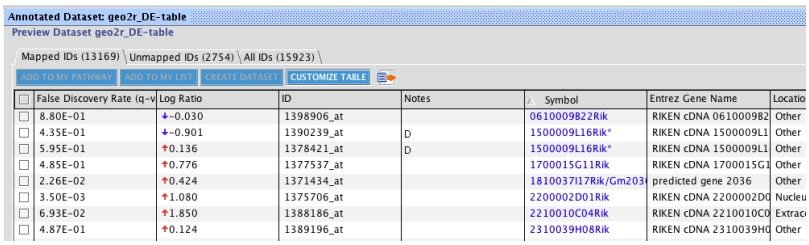
IPA Tutorial material

The starting material for the IPA upload can be downloaded from **this link (http://data.bits.vib.be/pub/trainingen/PUBMA2014/ex7-files/geo2r_DE-table.xls)** (bottom of this page)

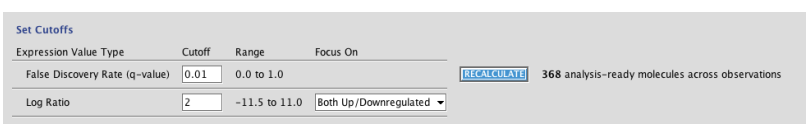
IPA analysis

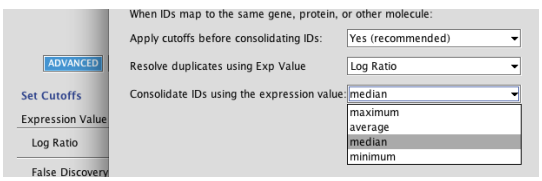
Upload data in IPA


Dataset Upload - geo2r_DE-table.xls
1. Select File Format:
Flexible Format
More Info
2. Contains Column Header:
Yes
No
3. Select Identifier Type:
Affymetrix
Specify the identifier type found in the dataset.
4. Array platform used for experiments:
Rat Expression Set 230 A
Select relevant array platform as a reference set for data analysis.
5. Use the dropdown menus to specify the columns that contain identifiers and observations. For observations, select the appropriate expression value type.
Raw Data (15924)
Dataset Summary (13169)
EDIT OBSERVATION NAMES
INFER OBSERVATIONS
More Info
ID
Observation 1
Ignore
Log Ratio
1 ID adj.P.Val P.Value t B logFC Gene.symbol Gene.title
2 1388876_at 7.67E-12 7.81E-16 59.1 25.28866 7.14
3 1374248_at 7.67E-12 9.78E-16 -57.9 25.14116 -11.5 Mybpc1 myosin binding p...
4 1374622_at 7.67E-12 1.44E-15 -56 24.87971 -4.69
5 1370033_at 3.47E-11 8.71E-15 -48 23.59223 -7.79 Myl1 myosin, light chai...
6 1371339_at 4.62E-11 1.45E-14 -46 23.20335 -8.25 Tnni1 troponin I type 1 ...
7 1373697_at 7.19E-11 2.71E-14 -43.6 22.71378 -7.13 Mybpc2 myosin binding p...
8 1398306_at 9.89E-11 4.35E-14 -41.8 22.33291 -7.06 Ampd1 adenosine mono...
9 1374672_at 1.14E-10 5.7E-14 40.8 22.11085 8.82 Tnni3k TNNI3 interacting...
10 1367962_at 1.44E-10 8.13E-14 -39.6 21.81731 -11.5 Actn3 actinin alpha 3


Annotated Dataset: geo2r_DE-table
Preview Dataset geo2r_DE-table
Mapped IDs (13169)
Unmapped IDs (2754)
All IDs (15923)
ADD TO MY PATHWAY
ADD TO MY LIST
CREATE DATASET
CUSTOMIZE TABLE
False Discovery Rate (q-v
Log Ratio
ID
Notes
Symbol
Entrez Gene Name
Locatio
8.80E-01 -0.030 1398906_at 0610009B22Rik RIKEN cDNA 061000982 Other
4.35E-01 -0.901 1390239_at D 1500009L16Rik* RIKEN cDNA 1500009L1 Other
5.95E-01 0.136 1378421_at D 1500009L16Rik* RIKEN cDNA 1500009L1 Other
4.85E-01 0.776 1377537_at 1700015G11Rik RIKEN cDNA 1700015G1 Other
2.26E-02 0.424 1371434_at 1810037I17Rik/Gm203 predicted gene 2036 Other
3.50E-03 1.080 1375706_at 2200002D01Rik RIKEN cDNA 2200002D0 Nucleu
6.93E-02 1.850 1388186_at 2210010C04Rik RIKEN cDNA 2210010C0 Extrac
4.87E-01 0.124 1389196_at 2310039H08Rik RIKEN cDNA 2310039H0 Other

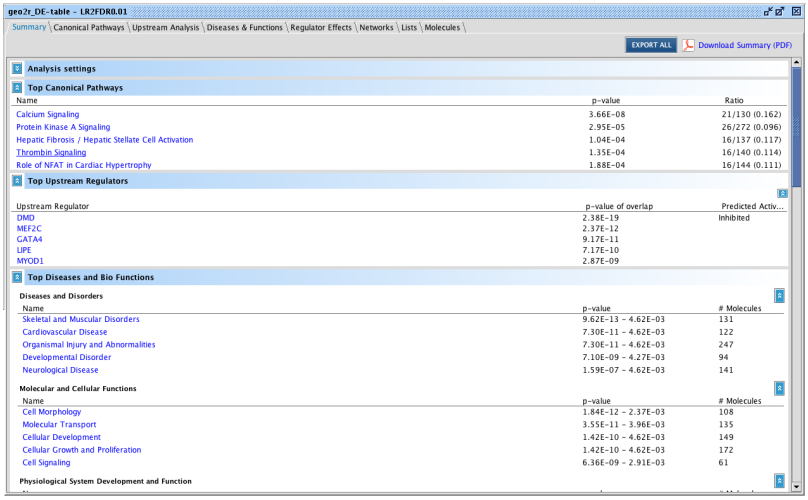
Start core analysis and set filter





Review the obtained core results

A full summary report can be downloaded from the server (see link at bottom of this page)


geo2r_DE-table - LR2FDR0.01
Summary \ Canonical Pathways \ Upstream Analysis \ Diseases & Functions \ Regulator Effects \ Networks \ Lists \ Molecules \
EXPORT ALL
Download Summary (PDF)
Analysis settings
Top Canonical Pathways
Name p-value Ratio
Calcium Signaling 3.66E-08 21/130 (0.162)
Protein Kinase A Signaling 2.95E-05 26/272 (0.096)
Hepatic Fibrosis / Hepatic Stellate Cell Activation 1.04E-04 16/137 (0.117)
Thrombin Signaling 1.35E-04 16/140 (0.114)
Role of NFAT in Cardiac Hypertrophy 1.88E-04 16/144 (0.111)
Top Upstream Regulators
Upstream Regulator p-value of overlap Predicted Activ...
DMD 2.38E-19 Inhibited
MEF2C 2.37E-12
GATA4 9.17E-11
LIPE 7.17E-10
MYOD1 2.87E-09
Top Diseases and Bio Functions
Diseases and Disorders
Name p-value # Molecules
Skeletal and Muscular Disorders 9.62E-13 - 4.62E-03 131
Cardiovascular Disease 7.30E-11 - 4.62E-03 122
Organismal Injury and Abnormalities 7.30E-11 - 4.62E-03 247
Developmental Disorder 7.10E-09 - 4.27E-03 94
Neurological Disease 1.59E-07 - 4.62E-03 141
Molecular and Cellular Functions
Name p-value # Molecules
Cell Morphology 1.84E-12 - 2.37E-03 108
Molecular Transport 3.55E-11 - 3.96E-03 135
Cellular Development 1.42E-10 - 4.62E-03 149
Cellular Growth and Proliferation 1.42E-10 - 4.62E-03 172
Cell Signaling 6.36E-09 - 2.91E-03 61
Physiological System Development and Function

Much more information and tools are available in IPA to continue the analysis and identify markers and targets for validation.

If you are a VIB scientist with 2-5 interested VIB colleagues, you may ask for a free custom 1-day training provided by BITS inside your lab. Please email us to discuss the possibilities

download exercise files

Download exercise files here. *[Expand]*

References:

[Main_Page | Hands-on Analysis of public microarray datasets | PubMA_Exercise.6]

Retrieved from "http://stelap.local/BioWareWIKI/index.php?title=PubMA_Exercise.7&oldid=11836"
Category: PUBMA2014

- This page was last modified on 21 October 2014, at 10:37.
- This page has been accessed 2 times.
- Content is available under Creative Commons Attribution Non-Commercial Share Alike unless otherwise noted.